

U.S.S.R. COMPUTATIONAL MATHEMATICS AND MATHEMATICAL PHYSICS

Cover-to-cover translation of
Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki

LIST OF CONTENTS AND AUTHOR INDEX
VOLUME 17, 1977



PERGAMON PRESS

OXFORD : NEW YORK : PARIS : FRANKFURT

Vol.
17
1977

U.S.S.R. COMPUTATIONAL MATHEMATICS AND MATHEMATICAL PHYSICS

Cover-to-cover translation of
Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki
Published bi-monthly by Pergamon Press

HONORARY EDITORIAL ADVISORY BOARD

R. A. BROOKER (<i>Manchester</i>)	A. S. HOUSEHOLDER (<i>Oak Ridge</i>)
L. FOX (<i>Oxford</i>)	G. F. J. TEMPLE (<i>Oxford</i>)
D. C. GILLES (<i>Glasgow</i>)	A. THOM (<i>Oxford</i>)
J. H. WILKINSON (<i>N.P.L., Teddington</i>)	

Scientific Translation Editor:
R. C. Glass, M.A., M.Sc.,
The City University, St. John Street, London, E.C.1.

Translators: J. Berry and D. E. Brown

Publishing, Subscription and Advertising Office:
Pergamon Press Ltd., Headington Hill Hall, Oxford, OX3 0BW

1979 Subscription rate \$300.00 (including postage and insurance)
Subscriptions are only serviced after payment has been received and
are not accepted for less than one year.

Orders should be sent to the Subscription Fulfillment Manager
Headington Hill Hall, Oxford, OX3 0BW

Microform Subscriptions and Back Issues

Back issues of all previously published volumes are available in the
regular editions and on microfilm and microfiche.

Current subscriptions are available on microfiche simultaneously with the
paper edition and on microfilm on completion of the annual index
at the end of the subscription year.

Copyright © 1978 Pergamon Press Ltd.

It is a condition of publication that manuscripts submitted to this journal have not been published and will not be simultaneously submitted or published elsewhere. By submitting a manuscript, the authors agree that the copyright for their article is transferred to the publisher if and when the article is accepted for publication. The copyright covers the exclusive rights to reproduce and distribute the article, including reprints, photographic reproductions, microform or any other reproductions of similar nature and translations. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, electrostatic, magnetic tape, mechanical, photocopying, recording or otherwise, without permission in writing from the copyright holder.

U.S. COPYRIGHT LAW APPLICABLE TO USERS IN THE U.S.A.

The Article Fee Code on the first page of an article in this journal indicates the copyright owner's consent that, in the U.S.A., copies may be made for personal or internal use provided the stated fee for copying, beyond that permitted by Section 107 or 108 of the United States Copyright Law, is paid. The appropriate remittance should be forwarded with a copy of the first page of the article to the Copyright Clearance Center Inc., PO Box 765, Schenectady, NY 12301. If a code does not appear copies of the article may be made without charge, provided permission is obtained from the publisher. The copyright owner's consent does not extend to copying for general distribution, for promotion, for creating new works or for resale. Specific written permission must be obtained from the publisher for such copying.

In case of doubt please contact your nearest Pergamon office.

PUBLISHED BY

PERGAMON PRESS LTD.

OXFORD · NEW YORK · PARIS · FRANKFURT

Vol
17
197

LIST OF CONTENTS

Number 1

A. I. GREBENNIKOV and V.A. MOROZOV	1	On the optimal approximation of operators
V. P. TANANA	12	On an iterative projection algorithm for solving ill posed problems with an approximately specified operator
L. T. POZNYAK	20	A new truncation procedure in the Bazely—Fox method
O. A. MALAFEEV	37	Stationary strategies in differential games
E. M. BERKOVICH	47	On a class of multi-stage problems of stochastic optimal control
N. S. VASIL'EV	58	An existence theorem in a minimax control problem
N. A. BOGOMOLOV and V. G. KARMANOV	65	A method of evaluating the stationary points of a general problem of non-linear programming
V. V. FEDOROV	72	Regularity conditions and necessary conditions for a maximin with connected variables
N. M. NOVIKOVA	84	A stochastic quasi-gradient method for seeking a maximin
V. I. LEBEDEV	92	An iterative method with Chebyshev parameters for finding the maximum eigenvalue and corresponding eigenfunction
YU. R. AKOPYAN and L. A. OGANESYAN	101	A variational-difference method for solving two-dimensional linear parabolic equations
Sh. M. NASIBOV	111	On numerical isolation of the bounded solutions of systems of linear partial differential equations of the evolutionary type
V. P. GORELOV and V. I. IL'IN	128	On rational approximation in Case's method
S. F. MOROZOV and I. P. SMIRNOV	140	Non-linear mathematical problems of the transmission of excitatory and inhibitory pulses in nerve tissue
A. D. KLIMOV, L. G. STRAKHOVSKAYA, R. P. FEDORENKO and I. L. CHIKHLADZE	154	The spatial kinetics of a pulsed heat-capacity reactor
R. Kh. ZEYTOUNIAN	167	Analyse asymptotique des écoulements de fluides visqueux compressibles à faible nombre de Mach
C. N. VOLOSHANOVSKAYA and M. M. KARCHEVSKII	174	A difference scheme for the problem of the strong bending of thin plates
L. V. SHIDLOVSKAYA	188	Numerical solution of the two-dimensional problem of shock wave propagation in outer space

V. I. MASLYANKIN	201	Convergence of the iterative process for the quasilinear heat transfer equation
A. V. KABULOV	210	Local algorithms on Yablonskii schemes
A. G. TSERKOVNYI	220	An approach to the construction of optimal recognition algorithms for large control tables
Short Communications		
A. YU. OSTROVSKII	233	The convergence of monotonic iterative processes
N. P. LIPATOV	233	A method of regularizing the inverse radon transformation in a medico-biological problem
V. A. GORDIN	237	Optimal quadrature formulas for a sphere
G. A. MIKHAILOV	244	Efficient Monte Carlo algorithms for evaluating the correlation characteristics of conditional mathematical expectations
V. YU. LEBEDEV	248	A search scheme for approximate solutions of the convex programming problem
V. D. NOGIN	254	Duality in multi-target programming
A. P. ABRAMOV and YU. P. IVANILOV	260	Algorithm for solving the linear programming problem by the loaded functional method
I. V. SIMONOV	265	The stability and asymptotic estimation of the solution of the inverse problem with a small parameter
A. L. GAPONENKO	270	A numerical method of solving three-dimensional diffraction problems
V. V. YANKOV	277	Solution of the inverse problem of the dispersion of an electromagnetic pulse in a conducting medium
YU. P. POPOV and E. A. SAMARSKAYA	281	Convergence of Newton's iterative method for solving gas-dynamic difference equations
	287	Book Review

Number 2

V. P. TANANA	1	On optimal methods for solving ill-posed problems with an approximately specified operator, and error estimates for the methods
G. P. GOLOVACH and A. F. KALAIDA	9	Solution of a spectral problem of transport theory by the splitting method
G. V. SAVINOV	17	Computational stability of matrix pivotal condensation
A. P. CHERENKOV	31	The distribution of diversified resources

E. A. AKHPATELOV and V. P. CHERENIN	38	Optimal serving sequence for a single system
A. B. VASIL'EVA and V. G. STEL'MAKH	48	Singularly disturbed systems of the theory of semiconductor devices
V. I. LEBEDEV	58	On a Zolotarev problem in the method of alternating directions
A. F. VOEVODIN	76	An implicit difference scheme for the integration of systems of hyperbolic equations
B. IOVANOVIICH	86	Additive difference scheme for a non-stationary fourth-order equation in an arbitrary domain
R. S. POPOVIDI and Z. S. TSVERIKMAZASHVILI	93	Numerical study of a diffraction problem by a modified method of non-orthogonal series
A. M. RADIN, V. A. REZUNENKO and V. P. SHESTOPALOV	104	Wave radiation by a sphere with a circular hole
F. A. GAREEV, S. A. GONCHAROV, E.P. ZHIDKOV, I.V. PUZYNIN, B.N. KHOROMSKII and R. M. YAMALEEV	116	Numerical solution of eigenvalue problems for nuclear-theory integro-differential equations
V. P. GORELOV and V. I. IL'IN	128	A correct asymptotic multigroup approximation of the problem of determining the critical radius of a sphere
M. I. VOLCHINSKAYA and B. N. CHETVERUSHKIN	135	An iterative method of solving the two-dimensional equations of radiative diffusion
V. G. KORNEEV and S. E. PONOMAREV	144	Solution of the plastic flow theory problem by the finite element method
L. S. KLABUKOVA and N. A. STADNIKOVA	161	Solution of mixed boundary value problems of the theory of momentless spherical shells by a differential—difference method
V. A. LYUL'KA	178	Numerical solution of the problem of the rotation of a cylinder in a flow of a viscous incompressible fluid
V. I. AVERTSEV and YU I. MOKIN	188	Algorithm for the control of grinding automata
J. LÖTZSCH	197	Experience of the computer realization of special languages by the depot system
Short Communications		
G. D. MAISTROVSKII	205	The conjugate gradient method in conditional minimization problems

B. S. PARIISKII	208	An economical method for the numerical solution of convolution equations
R. D. BAGLAI	211	The problem of identifying the functions generating a non-linear operator
V. M. VERBUCK and D. I. MILMAN	215	Veckstein's method as a modification of the transversal method
B. M. MUKHAMEDIEV	216	Construction of reachable sets for linear dynamic systems with noise
V. F. DEM'YANOV	220	An extremal basis method in minimax problems
I. G. MITEV	227	Application of the branch and bound method to some discrete programming problems
I. L. RUBANOV	232	Solution of the mixed boundary value problem for the Helmholtz equation on the exterior of a sphere
A. V. POPOV and S. A. KHOZIOSKII	238	A generalization of the parabolic equation of diffraction theory
A. I. EROFEEV and V. I. ZHUK	244	The scattering of a rarefied gas flow by a rough surface
L. G. SHUMKO	250	Analytic approximation of air pressure as a function of internal energy and density
	253	Book Reviews

Number 3

N. V. MUZYLEV	1	On the method of quasi-reversibility
G. S. GANSHIN	8	Optimal passive algorithms for evaluating the maximum of a function in an interval
S. M. ERMAKOV, A. I. PAVLOV and B. B. POKHODZEI	18	A method for evaluating multiple integrals with automatic choice of step
S. A. ABRAMOV	24	Second-order finite-difference equations with constant coefficients in the rational function field
A. I. KHISAMUTDINOV	29	Significance sampling and the simple Monte Carlo method for evaluating the sum of a Neumann series
G. SONNEVEND	35	On optimization of algorithms for function minimization
V. B. BISTRITSKAS	52	Optimal continuous analogues of multi-step optimal control processes
V. I. VENETS and M. V. RYBASHOV	64	The method of Lyapunov functions in the study of continuous algorithms of mathematical programming

A. B. YADYKIN	74	Parametrization in degenerate problems of quadratic programming
L. G. STRAKHOVSKAYA	88	An iterative method for evaluating the first eigenvalue of an elliptic operator
A. B. KUCHEROV and E. S. NIKOLAEV	101	An alternatively triangular iterative method for solving mesh elliptic equations in an arbitrary domain
N. N. KUZNETSOV	114	A finite difference method for solving the Cauchy problem for a quasi-linear first-order equation
S. I. SERDYUKOVA	127	Stability of boundary value problems for systems of difference equations with various structures
M. N. MURKES, V. A. ROZHDESTVENSKII and G. YU. SHOVRINSKII	133	Two numerical methods of solving one-dimensional problems of the filtering of multicomponent mixtures
S. A. GABOV	142	The angular potential and a problem with a directional derivative for harmonic functions
A. F. FILIPPOV	153	Integral representations of solutions of the two-dimensional wave equation and diffraction problems
YU. I. MOKIN	163	The mathematical model of the abrasion of a grinding wheel
V. P. KOROBEINIKOV, P. I. CHUSHKIN and L. V. SHURSHALOV	170	Allowance for atmospheric inhomogeneities in calculating the Tunguska meteorite explosion

Short Communications

A. M. DENISOV	184	Numerical solution of the converse scattering problem
FAM VAN AT	188	Isolation of the Gershgorin circles
M. V. ZAVOLZHENSKII and A. KH. TERSKOV	192	The zeros of the cylinder function $k_n(z)$
V. A. KANEVSKII and G. SH. LEV	196	Simulation of the point of emergence of Brownian motion onto the surface of a sphere
V. YU. LEBEDEV	198	Convergence of the weighted functional method in convex programming problems
WANG TAN-CHU	203	A modification of the Runge-Kutta procedure
S. L. ZIMONT and N. YA. MAR'YASHKIN	206	An analytic method of evaluating three-centre one-electron integrals with Slater functions
G. YA. SLEPYAN	211	Calculation of the natural electromagnetic oscillations of solids of revolution

O. I. LIETUVIETIS, G. A. RADZIN'SH and U. E. RAITUM	217	The optimization of plane-parallel magnetic fields
YU. D. SHMYGLEVSKII	222	A version of the moment method of calculating the transfer of selective radiation
S. O. BELOTSEKOVSKII and V. A. GUSHCHIN	228	Numerical simulation of the plane flow of a viscous fluid under the action of an external force periodic in space
	235	Book Reviews
Number 4		
	1	Nikita Nikolaevich Moiseev celebrates his Sixtieth birthday
E. L. ZHUKOVSKII	3	The method of least squares for degenerate and ill-posed systems of linear algebraic equations
HO THUAN	16	Some results concerning irreducible matrices
A. P. GRISHIN	26	The extremal problem of representing a given function of two variables by a table of values of functions of a single variable
V. V. VASIN	34	Order-wise optimality of the method of regularization for non-linear operator equations
A. I. PEROV and V. V. YURGELAS	45	On the convergence of an iterative process
M. L. AGRANOVSKII and R. D. BAGLAI	57	On an expansion in Hilbert space and its applications
A. A. LEVIKOV	64	Limit properties of a linear optimal control problem
YU. G. EVTUSHENKO and V. G. ZHADAN	73	A relaxation method of solving problems of non-linear programming
CHANG HAN	87	Approximate methods for solving problems of convex programming
V. V. FEDOROV	97	A method of solving linear hierarchical games
A. F. KONONENKO	104	On multi-step conflicts with information exchange
B. S. METEV	113	Game models of expert study
M. FRÖHNER	129	Spline solution of Cauchy problems for ordinary differential equations with delayed argument
YU. P. BOGLAEV	135	An iterative method for the approximate solution of differential-difference equations with a small retardation

G. F. ALIEV	150	Application of the method of integral relations and the method of quasi-linearization to Goursat's problem
G. P. ASTRAKHANTSEV	157	A method for the approximate solution of the Dirichlet problem for the biharmonic equation
V. P. GORELOV and V. I. IL'IN	175	Calculation of the critical radius of a shelled sphere
V. V. MOLOTKOV and E. I. URAZAKOV	185	The exact calculation of the excitation of some systems of plane wave-guides
N. B. MASLOVA	194	The solvability of stationary problems for Boltzmann's equation at large Knudsen numbers
Short Communications		
A. YU. AFANAS'EV and V. A. NOVIKOV	204	The search for a minimum of a function with a bounded third derivative
D. V. DENISOV	208	The method of coordinate descent in conditional minimization problems
G. P. AVANOVA	211	Singular controls in mathematical problems of physics
YU. YU. FINKEL'SHTEIN	215	The ϵ -approach to the multidimensional knapsack problem: the polynomial growth of the tree of branching
Y. G. MITEV	217	Some algorithms for solving non-linear integer mathematical programming problems
O. G. ALEKSEEV	222	A class of integer programming problems
M. G. VASIL'EV and V. S. YUFEREV	229	Approximation of discontinuous solutions of ordinary differential equations by polynomial splines
A. N. MINAILOS	235	Significance of the monotonicity of finite-difference schemes in shock-capturing methods
A. A. ARESEN'EV and N. V. PESKOV	241	On the existence of a generalized solution of Landau's equation
A. S. MEL'NICHENKO and V. N. OGIBIN	246	Application of the Monte Carlo method to the solution of spectral problems of radiative heat-exchange
A. A. AMOSOV, V. D. VALEDINSKII, YA. M. ZHILEIKIN and A. A. ZLOTNIK	253	Description of a set of programs for solving the light-wave propagation equations
V. M. KRIVTSOV, I. N. NAUMOVA, A. A. CHARAKHCH'YAN, A. V. SHIPILIN, YU. D. SHMYGLEVSKII and N. P. SHULISHNINA	256	Comparison of calculations of axisymmetric radiative gas flows
V. V. ARISTOV	261	The method of variable meshes in the velocity space in the of a strong condensation shock
	267	Book Reviews

Number 5

	1	Computer mathematics and scientific and technological progress
N. B. ENGIBARYAN and M. A. MNATSAKANYAN	3	Linear algebraic systems with Toeplitz matrices
F. GRUND	17	Inversion of five diagonal matrices
V. A. VARYUKHIN and S. A. KAS'YANYUK	23	On a class of iterative procedures for solving systems of non-linear equations
V. I. MELESHKO	32	Pseudo-inversion, stable to perturbations of closed operators
T. O. SHAPOSHNIKOVA	43	<i>A priori</i> error estimates for variational methods in Banach spaces
G. N. PLESHAKOV	51	On the efficiency of multi-dimensional interpolation iterations
G. V. KHROMOVA	58	Restoration of an inaccurate specified function
V. A. KARATYGIN and V. A. ROZOV	68	Asymptotic behaviour of sums with rapidly oscillating terms
V. A. ZHDANOV	80	On a method of coordinate descent
V. YU. LEBEDEV	86	A scheme for solving a partially integer-valued problem of mathematical programming
T. N. DANIL'CHENKO and K. K. MOSEVICH	91	A method of information transmission in multi-step two-person games
V. G. RAMM	100	On the importance of equilibrium in finite integral games
V. G. KORNEEV	109	Iterative methods of solving systems of equations of the finite element method
E. I. VELIEV and V. P. SHESTOPALOV	130	Diffraction of waves by a grating of circular cylinders with longitudinal slits
I. S. GUSHCHIN and YU. P. POPOV	144	Calculation of magnetohydrodynamic problems taking into account the phase transition
R. KH. ZEYTOUNIAN	152	Asymptotic analysis of the flows of viscous compressible fluids at low Mach numbers—II. The case of rotating heavy fluids
BAK HING KHANG	163	A parametric family of statistical recognition algorithms
R. G. NIGMATULLIN	174	The complexity of languages of type
T. L. SLUTSKAYA	181	Selection of the optimal algorithm in a class of recognition algorithms

Short Communications

M. M. SHUL'TS	192	The optimal bordering of matrices for Strassen's algorithm
O. M. MAKAROV	195	The lower bound of the number of multiplication operations for calculating the product of Hakei matrices
B. P. PUGACHEV	199	Acceleration of the convergence of iterative processes and a method of solving systems of non-linear equations
V. G. VASILEV	207	Estimates of some functionals by the solution of an integral equation of the first kind
S. B. OGNIVTSEV	210	A method of constructing attainability sets for non-linear control systems with phase constraints
M. G. RASSADINA and S. O. STRYGINA	215	On the convergence of the collocation method for non-linear boundary value problems
V. V. ZHAROVTSSEV	221	A completely conservative difference scheme for gas-dynamic equations
V. P. PARKHOMENKO, S. P. POPOV and O. S. RYZHOV	226	Effect of the initial particle velocity on the non-stationary spherically-symmetric motions of a gas

232 Book Review

Number 6

	1	Seventieth Birthday of Dmitrii Konstantinovich Faddeev
V. A. MOROZOV	3	Estimation of the accuracy of solving ill-posed problems and the solution of systems of linear algebraic equations
A. B. BAKUSHINSKII	12	Methods for solving monotonic variational inequalities based on the principle of iterative regularization
YU. G. DMITRIEV and V. V. KONEV	25	Error of the statistical estimation of multiple integrals using the ω^2 criterion
B. M. GOLUBITSKII and M. V. TANTASHEV	35	On local estimates in the Monte Carlo method
A. A. BELOLIPETSKII	40	A numerical method of solving a linear time optimal problem by reduction to a Cauchy problem
CHANG HAN	46	Some methods of constrained minimization in Hilbert space
R. P. FEDORENKO	54	Convergence of an iterative method for solving linear programming problems
V. D. SKARIN	65	Regularization of the min-max problems occurring in convex programming

L. G. KHACHIYAN	78	Convergence rate of the game processes for solving matrix games
V. G. PRIKAZCHIKOV	89	The finite difference eigenvalue problem for fourth-order elliptic operator
N. A. STRELKOV	100	On the choice of coordinate functions in projection-difference methods
V. F. BAKLANOVSKAYA	114	Study of the mesh method for parabolic equations with degeneration
G. K. KAISHIBAEVA and U. M. SULTANGAZIN	128	Construction of a difference scheme for systems of equations of the method of spherical harmonics
A. A. KIRILENKO, V. P. SHESTOPALOV and N. P. YASHINA	135	A rigorous solution of the problem of a circular waveguide with a step discontinuity of the cross-section
I. D. RODIONOV	146	Calculation of quantum-mechanical scattering by the isolated domain method
A. S. ALEKSEEV and A. G. MEGRABOV	158	Determination of the trajectory of motion of a pulse in the converse problem for the wave equation
L. M. DEGTYAREV and V. V. KRYLOV	172	A method for the numerical solution of problems of the dynamics of wave fields with singularities
A. A. GUBAIDULLIN, A. I. IVANDAEV and R. I. NIGMATULIN	180	A modified "coarse particle" method for calculating non-stationary wave processes in multiphase dispersive media
V. E. KARYAKIN and F. D. POPOV	193	Calculation of the three-dimensional supersonic flow of a viscous and heat-conducting gas past blunt bodies
A. N. NURLYBAEV	203	A local algorithm of index 1 for constructing the sum of dead-end disjunctive normal forms of functions of k -valued logic
A. I. ZENKIN and V. V. RYAZANOV	211	Algorithms predicting the states of controlled systems

Short Communications

O. M. MAKAROV	221	An algorithm for the multiplication of two quaternions
N. S. STADNIKOVA	222	An auxiliary algebraic problem
T. I. SAVELOVA	227	The optimal regularization of operator equations with errors in the definition of the operators and the right side
A. T. GASHIMOV and S. Ya. YAKUBOV	232	A finite-difference method of solving the Cauchy problem for evolution equations
A. S. ZIL'BERGLEIT	237	A uniform asymptotic expansion of Dirichlet's integral
L. A. ISTOMIN	242	A modification of Hoang T'ui's method of minimizing a concave function on a polyhedron

I. G. MITEV

N. N. ELKIN and V. V.
KRAVTSOV

S. P. POPOV and Yu. I.
ROMASHKEVICH

248 On a penalty in discrete programming

250 Numerical study of the problem of wave diffraction in a
rotating basin

254 Application of the splitting method for calculating two-
temperature and ionizationally non-equilibrium gas flows

264 Book Reviews

Vol.
17
1977

AUTHOR INDEX

ABRAMOV, A. P. 260 (1)*
 ABRAMOV, S. A. 24 (3)
 AFANAS'EV, A. Yu. 204 (4)*
 AGRANOVSKII, M. L. 57 (4)
 AKHPATELOV, E. A. 38 (2)
 AKOPYAN, Yu. R. 101 (1)
 ALEKSEEV, A. S. 158 (6)
 ALEKSEEV, O. G. 222 (4)*
 ALIEV, G. F. 150 (4)
 AMOSOV, A. A. 253 (4)*
 ARISTOV, V. V. 261 (4)*
 ARSEN'EV, A. A. 241 (4)*
 ASTRAKHANTSEV, G. P. 157 (4)
 AVERTSEV, V. I. 188 (2)
 AVANOVA, G. P. 211 (4)*

 BAGLAI, R. D. 211 (2),* 57 (4)
 BAKLANOVSKAYA, V. F. 114 (6)
 BAKUSHINSKII, A. B. 12 (6)
 BELOLIPETSKII, A. A. 40 (6)
 BELOTSEKOVSKII, S. O. 228 (3)*
 BERKOVICH, E. M. 47 (1)
 BISTRITSKAS, V. B. 52 (3)
 BOGLAEV, Yu. P. 135 (4)
 BOGOMOLOV, N. A. 65 (1)

 CHARAKHCH'YAN, A. A. 256 (4)*
 CHERENIN, V. P. 38 (2)
 CHERENKOV, A. P. 31 (2)
 CHETVERUSHKIN, B. N. 135 (2)
 CHIKHLADZE, I. L. 154 (1)
 CHUSHKIN, P. I. 170 (3)

 DANIL'CHENKO, T. N. 91 (5)
 DEGTYAREV, L. M. 172 (6)
 DEM'YANOV, V. F. 220 (2)*
 DENISON, A. M. 184 (3)*
 DENISOV, D. V. 208 (4)*
 DMITRIEV, Yu. G. 25 (6)

 ELKIN, N. N. 250 (6)*
 ENGIBARYAN, N. B. 3 (5)
 ERMAKOV, S. M. 18 (3)
 EROFEEV, A. I. 244 (2)*
 EVTUSHENKO, Yu. G. 73 (4)

 FEDORENKO, R. P. 154 (1), 54 (6)
 FEDOROV, V. V. 72 (1), 97 (4)

FILIPPOV, A. F. 153 (3)
 FINKEL'SHTEIN, Yu. Yu. 215 (4)*
 FROHNER, M. 129 (4)

 GABOV, S. A. 142 (3)
 GANSHIN, G. S. 8 (3)
 GAPONENKO, A. L. 270 (1)*
 GAREEN, F. A. 116 (2)
 GASHIMOV, A. T. 232 (6)*
 GOLOVACH, G. P. 9 (2)
 GOLUBITSKII, B. M. 35 (6)
 GONCHAROV, S. A. 116 (2)
 GORDIN, V. A. 237 (1)*
 GORELOV, V. P. 128 (1), 128 (2), 175 (4)
 GREBENNIKOV, A. I. 1 (1)
 GRISHIN, A. P. 26 (4)
 GRUND, F. 17 (5)
 GUBAIDULLIN, A. A. 180 (6)
 GUSHCHIN, I. S. 144 (5)
 GUSHCHIN, V. A. 228 (3)*

 HAN, C. 87 (4), 46 (6)

 IL'IN, V. I. 128 (1), 128 (2), 175 (4)
 IOVANOVICH, B. 86 (2)
 ISTOMIN, L. A. 242 (6)*
 IVANDAEV, A. I. 180 (6)

 KAISHIBAEVA, G. K. 128 (6)
 KALAIDA, A. F. 9 (2)
 KANEVSKII, V. A. 196 (3)*
 KARATYGIN, V. A. 68 (5)
 KARCHEVSKII, M. M. 174 (1)
 KABULOV, A. V. 210 (1)
 KARMANOV, V. G. 65 (1)
 KARYAKIN, V. E. 193 (6)
 KAS'YANUK, S. A. 23 (5)
 KHACHIYANZ, L. G. 78 (6)
 KHANG, B. H. 163 (5)
 KHISAMUTDINOV, A. I. 29 (3)
 KHOROMSKII, B. N. 116 (2)
 KHOZIOSKII, S. A. 238 (2)*
 KHROMOVA, G. V. 58 (5)
 KIRILENKO, A. A. 135 (6)
 KLABULOVA, L. S. 161 (2)
 KLIMOV, A. D. 154 (1)
 KONEV, V. V. 25 (6)
 KONONENKO, A. F. 104 (4)
 KORNEEV, V. G. 144 (2), 109 (5)
 KOROBEINIKOV, V. P. 170 (3)
 KRAVTSOV, V. V. 250 (6)*
 KRIVTSOV, V. M. 256 (4)*

*Short communications

- KRYLOV, V. V. 172 (6)
 KUCHEROV, A. B. 101 (3)
 KUZNETSOV, N. N. 114 (3)
- LEBEDEV, V. I. 92 (1), 58 (2)
 LEBEDEV, V. Yu. 248 (1)*, 198 (3)*, 86 (5)
 LEV, G. Sh. 196 (3)*
 LEVIKOV, A. A. 64 (4)
 LIETUVIETIS, O. I. 217 (3)*
 LIPATOV, N. P. 233 (1)*
 LOTZSCH, J. 197 (2)
 LYUL'KA, V. A. 178 (2)
 MAISTROVSKII, G. D. 205 (2)*
 MAKAROV, O. M. 195 (5)*, 221 (6)*
 MALAFEEV, O. A. 37 (1)
 MAR'YASHKIN, N. YA. 206 (3)*
 MASLOVA, N. B. 194 (4)
 MASLYANKIN, V. I. 201 (1)
 MEGRABOV, A. G. 158 (6)
 MELESHKO, V. I. 32 (5)
 MEL'NICHENKO, A. S. 246 (4)*
 METEV, B. S. 113 (4)
 MIKHAILOV, G. A. 244 (1)*
 MILMAN, D. I. 215 (2)*
 MINAILOS, A. N. 235 (4)*
 MITEV, I. G. 248 (6)*
 MITEV, I. G. 227 (2)*
 MITEV, Y. G. 217 (4)*
 MNATSAKANYAN, M. A. 3 (5)
 MOKIN, Yu. I. 188 (2), 163 (3)
 MOLOTKOV, V. V. 185 (4)
 MOROZOV, S. F. 140 (1)
 MOROZOV, V. A. 1 (1), 3 (6)
 MOSEVICH, K. K. 91 (5)
 MUKHAMEDIEV, B. M. 216 (2)*
 MURKES, M. N. 133 (3)
 MUZYLEV, N. V. 1 (3)
- NASIBOV, SH. M. 111 (1)
 NAUMOVA, I. N. 256 (4)*
 NIGMATULIN, R. I. 180 (6)
 NIGMATULLIN, R. G. 174 (5)
 NIKOLAEV, E. S. 101 (3)
 NOGIN, V. D. 254 (1)*
 NOVIKOV, V. A. 204 (4)*
 NOVIKOVA, N. M. 84 (1)
 NURLYBAEV, A. N. 203 (6)
- OGANESYAN, L. A. 101 (1)
 OGIBIN, V. N. 246 (4)*
 OGNIVTSEV, S. B. 210 (5)*
 OSTROVSKII, A. Yu. 227 (1)*
- PARIISKII, B. S. 208 (2)*
 PARKHOMENKO, V. P. 226 (5)*
- PAVLOV, A. I. 18 (3)
 PEROV, A. I. 45 (4)
 PESKOV, N. V. 241 (4)*
 PLESHAKOV, G. N. 51 (5)
 POKHODZEI, B. B. 18 (3)
 PONOMAREV, S. E. 144 (2)
 POPOV, F. D. 193 (6)
 POPOV, S. P. 226 (5)*, 254 (6)*
 POPOV, Yu. P. 281 (1)*, 144 (5)
 POPOVIDI, R. S. 93 (2)
 POPOV, A. V. 238 (2)*
 POZNYAK, L. T. 20 (1)
 PRIKAZCHIKOV, V. G. 89 (6)
 PUGACHEV, B. P. 199 (5)*
 PUZYNNIN, I. V. 166 (2)
- RADIN, A. M. 104 (2)
 RADZIN'SH, G. A. 217 (3)*
 RAITUM, V. E. 217 (3)*
 RAMM, V. G. 100 (5)
 REZUNENKO, V. A. 104 (2)
 RASSADINA, M. G. 215 (5)*
 RODIONOV, I. D. 146 (6)
 ROMASHKEVICH, Yu. I. 254 (6)*
 ROZHDESTVENSKII, V. A. 133 (3)
 ROZOV, V. A. 68 (5)
 RUBANOV, I. L. 232 (2)*
 RYAZANOV, V. V. 211 (6)
 RYBASHOV, M. V. 64 (3)
 RYZHOV, O. S. 226 (5)*
- SAMARSKAYA, E. A. 281 (1)*
 SAVELOVA, T. I. 227 (6)*
 SAVINOV, G. V. 17 (2)
 SERDYUKOVA, S. I. 127 (3)
 SHAPOSHNIKOVA, T. O. 43 (5)
 SHESTOPALOV, V. P. 104 (2), 130 (5), 135 (6)
 SHIDLOVSKAYA, L. V. 188 (1)
 SHIPILIN, A. V. 256 (4)*
 SHMYGLEVSKII, Yu. D. 222 (3)*, 256 (4)
 SHOVRKINSKII, G. Yu. 133 (3)
 SHULISHNINA, N. P. 256 (4)*
 SHUL'TS, M. M. 192 (5)*
 SHUMKO, L. G. 250 (2)*
 SHURSHALOV, L. V. 170 (3)
 SIMONOV, I. V. 265 (1)*
 SLUTSKAYA, T. L. 181 (5)
 SKARIN, V. D. 65 (6)
 SLEPYAN, G. Ya. 211 (3)*
 SMIRNOV, I. P. 140 (1)
 SONNEVEND, G. 35 (3)
 STADNIKOVA, N. A. 161 (2), 222 (6)*
 STEL'MAKH, V. G. 48 (2)
 STRAKHOVSKAYA, L. G. 154 (1), 88 (3)
 STRELKOV, N. A. 100 (6)

STRYGINA, S. O. 215 (5)*
SULTANGAZIN, U. M. 128 (6)

TANANA, V. P. 12 (1), 1 (2)
TAN-CHU, W. 203 (3)*
TANTASHEV, M. V. 35 (6)
TERSKOV, A. Kh. 192 (3)*
THUAN, H. 16 (4)
TSVERIKMAZASHVILI, Z. S. 93 (2)
TSERKOVNYI, A. G. 220 (1)
URAZAKOV, E. I. 185 (4)

VALEDINSKII, V. D. 253 (4)*
VAN AT, F. 188 (3)*
VARYUKHIN, V. A. 23 (5)
VASIL'EV, M. G. 229 (4)*
VASIL'EV, N. S. 58 (1)
VASILEV, V. G. 207 (5)*
VASIL'EVA, A. B. 48 (2)
VASIN, V. V. 34 (4)
VELIEV, E. I. 130 (5)
VENETS, V. I. 64 (3)
VERBUCK, V. M. 215 (2)*
VOEVODIN, A. F. 76 (2)

VOLCHINSKAYA, M. I. 135 (2)
VOLOSHANOVSKAYA, C. N. 174 (1)

YADYKIN, A. B. 74 (3)
YAKUBOV, S. Ya. 232 (6)*
YAMALEEV, R. M. 116 (2)
YANKOV, V. V. 277 (1)*
YASHINA, N. P. 135 (6)
YUFEREV, V. S. 229 (4)*
YURGELAS, V. V. 45 (4)

ZAVOLZHENSKII, M. V. 192 (3)*
ZENKIN, A. I. 211 (6)
ZEYTOUNIAN, R. Kh. 167 (1), 152 (5)
ZHADAN, V. G. 73 (4)
ZHAROVTSSEV, V. V. 221 (5)*
ZHDANOV, V. A. 80 (5)
ZHIDKOV, E. P. 116 (2)
ZHILEIKIN, Ya. M. 253 (4)*
ZHUK, V. I. 244 (2)*
ZHUKOVSKII, E. L. 3 (4)
ZIL'BERGLEIT, A. S. 237 (6)*
ZIMONT, S. L. 206 (3)*
ZLOTNIK, A. A. 253 (4)*

ON THE OPTIMAL APPROXIMATION OF OPERATORS*

A. I. GREBENNIKOV and V. A. MOROZOV

Moscow

(Received 31 March 1975)

WE POSE the problem of approximating optimally the values of an unbounded operator, using elements which are specified solely by the traces (values) of some operator. Estimates are obtained for the accuracy of the optimal approximation, linear optimal algorithms are found in explicit form, and their structures are examined, and the optimal algorithm is shown to be unique. Some examples are given.

Many problems in the processing of experimental data can be formulated as a problem in computing the values of an unbounded operator. In general, this problem is ill posed [1]. Often, the initial information contains errors, and arrives in discrete form, with the result that the value of the operator can only be evaluated approximately. It thus becomes extremely important to find the (in some sense) best or optimal operator, approximating the initial operator.

The problem of the optimal approximation of an operator when the information is exactly specified was considered by Stechkin in [2] and Bakhvalov in [3]. The optimal approximation of a linear bounded functional when the information is specified approximately was considered by Marchuk and Osipenko in [4], and by Reinsch in [5]. A particular case of the approximation of operators was studied in [6]. Order-wise optimal linear operators were found by Morozov in [7]. Mention may also be made of papers by Strakhov [8], by Ivanov and Korolyuk [9], and by V. V. Ivanov [10].

Let us state our problem. Let H, G, F, V be normed spaces. Let L be a linear operator with non-empty domain of definition $D_L \subset H$, mapping D_L into the space G , and let A be a linear operator with domain of definition $D_A \subset H$, mapping D_A into F . We shall assume that $D = D_L \cap D_A \neq \emptyset$. Finally, let B be the operator to be evaluated, with domain of definition $D_B \subset H$, mapping D_B into V , such that $D \subseteq D_B$. The operators A, B, L may be unbounded; here, $B(-u) = -Bu \forall u \in D_B$.

We are given the admissible set of elements

$$M_R = \{u \in D : \|Lu\|_G \leq R\}, \quad 0 < R = \text{const} < +\infty.$$

On the basis of information about the element $u \in M_R$, characterizing the exact (or approximate) value of the operator A on the element: $f = Au$ or $f \approx Au$, we have to compute the value of the operator Bu .

The cases of both exact and approximate specification of the element f will be considered. Assuming the existence of supplementary *a priori* information about the element u , we find an effective lower bound for the error of the approximation, which is independent of the approximating operator. In the case of a linear operator B and Hilbert spaces H, F , and G , we find

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 3-14, 1977.

the approximation, best in the sense of some chosen criterion, to the operator B , both at every point f (or \tilde{f}) and in the entire set of data. In the case of approximation at a point, the optimal operator is shown to be unique.

The existence of the optimal operator is proved by a functional method and is not based on geometric considerations as e.g., in [3]. Another important point is that the optimal operator itself is found during the proof of existence; this operator proves to be linear. Moreover, regardless of the concrete form of the operator B , the optimal operator has the following structure: $T_{\text{opt}} = B \cdot S$, where S is independent of the operator B , and is fully defined by the initial data of the problem. This means that operators which are optimal in a class of problems can be constructed.

1. Exactly specified information

1. Assume that the information about the element u is specified exactly. We introduce the set of all data N_R :

$$N_R = \{f \in F : f = Au, u \in M_R\}.$$

The set $M_R \neq \emptyset$, so that N_R also is non-empty.

Given the fixed element $f \in N_R$. We introduce the set

$$U_R(f) = \{u \in M_R : Au = f\}.$$

Concerning the element u , on which we want to find the value of the operator B , assume that it is known *a priori* that $u \in U_R(f)$. Denote by T any operator (not necessarily linear) which is defined in N_R and maps N_R into V . The error of approximation of the operator B by means of the operator T on the set N_R will be characterized by the function

$$\omega_B(R, T) = \sup_{f \in N_R, u \in U_R(f)} \|Bu - Tf\|_V.$$

We put

$$\omega_B(R) = \sup \|Bu\|_V, \quad u \in U_R = \{u \in M_R : Au = 0\}.$$

It is assumed that $\omega_B(R)$ is defined (finite) for all $R > 0$. For this, it is sufficient that the operator B satisfy the B -complementarity condition [7]

$$\|Bu\|_V^2 \leq \gamma_B (\|Au\|_F^2 + \|Lu\|_G^2) \quad \forall u \in D, \quad 0 < \gamma_B = \text{const} < +\infty.$$

We have:

Theorem 1

Under the above assumptions, for any admissible operator T we have the lower bound

$$\omega_B(R, T) \geq \omega_B(R). \quad (1)$$

Proof. Let $h_\varepsilon \in U_R$ be an element such that

$$\|Bh_\varepsilon\|_V \geq \omega_B(R) - \varepsilon,$$

where $\epsilon > 0$ is an arbitrary number. It is obvious that the element $(-h_\epsilon) \in U_R$, and since the operator B is homogeneous, $\omega_B(R) \leq \|B(-h_\epsilon)\|_V + \epsilon$. By the definition of $\omega_B(R, T)$, we have

$$\omega_B(R, T) \geq \max\{\|Bh_\epsilon - T\Theta\|_V, \|Bh_\epsilon + T\Theta\|_V\},$$

where Θ is the zero of the space F . Since

$$2Bh_\epsilon = Bh_\epsilon - T\Theta + (Bh_\epsilon + T\Theta),$$

we obtain, using the triangle inequality,

$$\begin{aligned} 2\|Bh_\epsilon\|_V &\leq \|Bh_\epsilon - T\Theta\|_V + \|Bh_\epsilon + T\Theta\|_V \\ &\leq 2 \max\{\|Bh_\epsilon - T\Theta\|_V, \|Bh_\epsilon + T\Theta\|_V\}. \end{aligned}$$

Hence

$$\omega_B(R, T) \geq \|Bh_\epsilon\|_V \geq \omega_B(R) - \epsilon.$$

Recalling that ϵ is arbitrary, the theorem now follows.

Notes. 1. If the kernel of the operator $A: N_A = \{u \in D: Au = 0\}$ consists solely of zero, i.e., A is invertible, then $U_R = \{0\}$. Then, obviously, $\omega_B(R) = 0 \quad \forall R > 0$. It is therefore natural to require that $\dim N_A \geq 1$.

2. The same device was used in the proof of Theorem 1 as was used for the proof of a similar theorem in [4]; see also [7].

The lower bound (1) has been proved for a non-linear operator B . It will be shown below that the bound can be reached in the case of a linear operator B .

2. We introduce the error of approximation of the operator B by an operator T at the point f :

$$\omega_B(R, T, f) = \sup_{u \in U_{R(f)}} \|Bu - Tf\|_V.$$

Consider the following problems of optimal approximation of the operator B : to find the operator T_0 , optimal at a point, i.e., such that

$$\omega_B(R, T_0, f) = \inf_T \omega_B(R, T, f) = \omega_B(R, f); \quad (2)$$

and to find the operator P_0 , optimal in N_R , i.e., such that

$$\omega_B(R, P_0) = \inf \omega_B(R, T) = \omega_B(R, N_R). \quad (3)$$

We shall assume now that H, F , and G are Hilbert spaces. For fixed $f \in N_R$ we introduce the set

$$U(f) = \{u \in D: Au = f\}.$$

To solve problems (2) and (3), we consider an auxiliary problem: to find the element $u_f \in D$ such that

$$\|Lu_f\|_G = \inf \|Lu\|_G, \quad u \in U(f); \quad (4)$$

We shall say that the operators A and L are jointly closed in D if, given any sequence of elements $u_n \in D$, such that

$$\lim_{n \rightarrow \infty} u_n = u_0 \quad (B H), \quad \lim_{n \rightarrow \infty} Au_n = f_0 \quad (B F), \quad \lim_{n \rightarrow \infty} Lu_n = g_0 \quad (B G),$$

it follows that $u_0 \in D$, and $Au_0 = f_0$, $Lu_0 = g_0$.

It was shown in [7] that, provided the operators be jointly closed, an element $u_f \in D$, which we shall call the L -pseudo-solution, exists and is unique for every $f \in A[D]$, provided that, for any $u \in D$, we have the complementarity condition

$$\|u\|_H^2 \leq \gamma (\|Au\|_F^2 + \|Lu\|_G^2) = \gamma \|u\|_{A,L}^2, \quad 0 < \gamma = \text{const} < +\infty.$$

A linear operator $S_0 : S_0 f = u_f$ is thereby defined, with domain of definition $D_{S_0} = A[D]$.

Notice that the equation $Au = f$ is solvable for any $f \in F$, if $Q_A = \{f \in F : f = Au, u \in D\} = F$. In this case, the operator S_0 is defined in the whole of F , and hence is bounded in D , equipped with the norm $\|u\|_{A,L}$.

Lemma 1

For all $u \in U(f)$ we have

$$\|Lu_f - Lu\|_G \leq \|Lu\|_G. \quad (5)$$

Proof. From Euler's identity for problem (4) we obtain, for all $u \in U(f)$;

$$(Lu_f - Lu, Lu_f)_G = 0. \quad (6)$$

From the obvious identity

$$\|Lu_f - Lu\|_G^2 = \|Lu\|_G^2 - 2(Lu - Lu_f, Lu_f)_G + \|Lu_f\|_G^2$$

and Eq. (6), we obtain, for any $u \in U(f)$, the Pythagoras equation

$$\|Lu\|_G^2 = \|Lu_f\|_G^2 + \|Lu - Lu_f\|_G^2. \quad (7)$$

The inequality (5) follows obviously from (7). The lemma is proved.

Note 3. If $f = Au$, $u \in N_L$, where the set $N_L = \{u \in D : Lu = 0\}$, then it can easily be seen that we have the relation $u_f = u \quad \forall f$.

Theorem 2

If B is a linear operator, then the same linear operator is a solution of problems (2) and (3), namely,

$$T_0 = BS_0,$$

the absolute values of the errors being respectively

$$\omega_B(R, f) = \sup \|Bu - BS_0 f\|_V, \quad u \in U_R(f), \quad (8)$$

$$\omega_B(R, N_R) = \omega_B(R). \quad (9)$$

If V is a Hilbert space, then T_0 is the unique solution of problem (2).

Proof. For the pseudo-solution u_f and arbitrary $u \in U(f)$ we have

$$\|L(2u_f - u)\|_G = \|Lu\|_G.$$

Hence, for any $u \in U_R(f)$, the element $(2u_f - u)$ also lies in $U_R(f)$. Moreover, for any operator T , any $u \in U_R(f)$ and any $f \in N_R$ we have the obvious inequality

$$\omega_B(R, T, f) \geq \max\{\|B(2u_f - u) - Tf\|_V, \|Bu - Tf\|_V\}. \quad (10)$$

From the equation

$$2B(u_f - u) = B(2u_f - u) - Tf + (Tf - Bu),$$

which holds by virtue of the fact that the operator B is linear, we obtain with the aid of the triangle inequality

$$2\|Bu_f - Bu\|_V \leq 2\max\{\|B(2u_f - u) - Tf\|_V, \|Bu - Tf\|_V\}.$$

We then obtain from (10):

$$\omega_B(R, T, f) \geq \sup_u \|Bu - BS_0 f\|_V, \quad u \in U_R(f). \quad (11)$$

Obviously, the equality is reached in (11) for $F = BS_0$, whence it follows that T_0 is the solution of problem (2), and Eq. (8) holds.

If $u \in U_R(f)$, then, by the properties of the pseudo-solution, we have $A(u - u_f) = 0$. From (5) we obtain

$$\|L(u - u_f)\|_G \leq R.$$

Consequently, the element $(u - u_f) \in U_R$ for any $u \in U_R(f)$ and any $f \in N_R$. Hence, for any $f \in N_R$,

$$\omega_B(R) = \sup_{u \in U_R} \|Bu\|_V \geq \sup_{u \in U_R(f)} \|Bu - Bu_f\|_V.$$

It then follows from (1) that, for any T ,

$$\omega_B(R, T) \geq \omega_B(R) \geq \sup_{f \in N_R} \sup_{u \in U_R(f)} \|Bu - BS_0 f\|_V. \quad (12)$$

The extreme terms in (12) are obviously the same for $T = T_0$, and hence T_0 is also a solution of problem (3). Equation (9) follows obviously from (12).

Now let V be a Hilbert space. We shall prove that the operator T_0 , optimal at a point, is unique. Let \tilde{T} be another optimal operator. We shall find the operator $\tilde{T}_{1/2}$ such that

$$\tilde{T}_{1/2} f = \frac{1}{2} \tilde{T} f + \frac{1}{2} BS_0 f, \quad f \in N_R.$$

The operator $\tilde{T}_{1/2}$ is also optimal at a point, since

$$\sup_{u \in U_R(f)} \|Bu - \tilde{T}_{1/2} f\|_V \leq^{1/2} \sup_{u \in U_R(f)} \|Bu - \tilde{T} f\|_V +^{1/2} \sup_{u \in U_R(f)} \|Bu - BS_0 f\|_V = \omega_B(R, f).$$

On applying the parallelogram equation and recalling that the operators \tilde{T} , BS_0 , $\tilde{T}_{1/2}$, are optimal, we obtain for arbitrary T the inequality

$$\omega_B^2(R, T, f) \geq \omega_B^2(R, f) + 1/8 \|BS_0 f - Tf\|_V^2.$$

Hence it is clear that, to reach the optimal error $\omega_B(R, f)$, it is necessary that

$$\|BS_0 f - Tf\|_V = 0.$$

This relation proves the uniqueness.

Note 4. Using (8) and Note 3, we get

$$\omega_B(R, f) = 0 \quad \forall f: f = Au, \quad u \in N_L,$$

i.e., on elements of the kernel of the operator L , the algorithm T_0 gives exact values for any admissible operator B .

Given the linear operator B_0 in D_B , acting from H into V , such that

$$c_1 \|B_0 u\|_V \leq \|Bu\|_V \leq c_2 \|B_0 u\|_V,$$

where $c_i > 0$ are constants, independent of $u \in D_B$. We then have:

Lemma 2. Under the conditions stated, we have the estimates

$$c_1 \omega_B^0(R) \leq \omega_B(R) \leq c_2 \omega_B^0(R),$$

where

$$\omega_B^0(R) = \sup_u \|B_0 u\|_V, \quad u \in U_R.$$

The proof is obvious.

Lemma 2 shows that the function $\omega_B^0(R)$ has the same order in R as the function $\omega_B(R)$, though the evaluation of it, when finding the order of accuracy of the approximation of the operator B , can sometimes prove to be far simpler.

3. Notice that the optimal operator T_0 is independent of the number R ; only the errors $\omega_B(R, f)$ and $\omega_B(R, N_R)$ depend on R . We shall consider as a measure of the accuracy of approximation of the operator B the quantities

$$\hat{\omega}_B(T, f) = \sup_u \{\|Bu - Tf\| / \|Lu\|\}, \quad u \in U(f), \quad \|Lu\| \neq 0,$$

$$\hat{\omega}_B(T) = \sup_f \hat{\omega}_B(T, f), \quad f \in A[D].$$

We shall seek the operators \hat{T}_0 and \hat{P}_0 such that

$$\hat{\omega}_B(\hat{T}_0, f) = \inf_T \hat{\omega}_B(T, f) = \hat{\omega}_B(f), \quad (13)$$

$$\hat{\omega}_B(\hat{P}_0) = \inf_T \hat{\omega}_B(T) = \hat{\omega}_B. \quad (14)$$

The proof of the following theorem is similar to the proof of Theorem 2.

Theorem 3

If B is a linear operator, then the same linear operator

$$T_0 = \hat{T}_0 = \hat{P}_0,$$

provides a solution of problems (13) and (14); here,

$$\hat{\omega}_B(f) = \sup_u \{ \|Bu - BS_0 f\| / \|Lu\| \}, \quad u \in U(f), \quad \|Lu\| \neq 0,$$

$$\hat{\omega}_B = \sup_f \hat{\omega}_B(f), \quad f \in A[D].$$

4. Let us give some examples. Let $D = D_L$ and

$$Au = (A_1 u, \dots, A_n u),$$

where A_i is a linear bounded operator, defined in D and mapping the Hilbert space H into the Hilbert space F_i . We define the space F as the Cartesian product of the spaces F_i :

$$F = F_1 \times F_2 \times \dots \times F_n,$$

with the norm $\|f\|_F^2 = \|f_1\|_{F_1}^2 + \dots + \|f_n\|_{F_n}^2$, $f = (f_1, \dots, f_n) \in F$.

It follows from Theorem 2 that the method of operator (and also functional) interpolational splines (see [7], p. 278) is optimal for evaluating the values of a linear operator B , given the *a priori* information $u \in U_R(f)$, $f \in N_R$.

It can be shown in a similar way that the modified method of collocation, see [11, 12], is also optimal. Consider some simple examples:

Example 1. Let $H = W_2^1[0, 1]$, $G = L_2[0, 1]$, $F = R_n$, $V = C[0, 1]$, $L = d/dx$, $A_i u(x) = u(x_i)$, $x_i = ih$, $i = 0, 1, \dots, n$, $h = 1/n$, $B = E$. Assume that it is known that

$$u \in M_R = \left\{ u : \int_0^1 [u'(x)]^2 dx \leq R^2 \right\}.$$

This case corresponds to finding the method of best uniform approximation of the function $u(x)$ on the set M_R .

By Theorem 2, the best uniform approximation is the interpolational spline of the first degree (step-line).

Example 2. Let $H = W_2^2[0, 1]$, $G = L_2[0, 1]$, $F = R_n$, $V = C[0, 1]$, $L = d^2/dx^2$, $A_i u(x) = u(x_i)$, $x_i = ih$, $i = 0, 1, \dots, n$, $h = 1/n$. $B = d/dx$, i.e., we consider the problem of best uniform approximation of the first derivative of the function $u(x)$. Assume that it is known that

$$u \in M_R = \left\{ u : \int_0^1 [u''(x)]^2 dx \leq R^2 \right\}.$$

By Theorem 2, the best approximation of $u'(x)$ is the function $s'(x)$, where $s(x)$ is the cubic interpolational spline.

The results described above can be extended in a natural way to the case of unsolvable equations $Au = f$. In this case, the element f has to be replaced by the element Pf , where P is the operator of orthogonal projection onto the set Q_A , if $Pf \in Q_A$ (we have in mind the case when the closure of Q_A in F is not the same as Q_A).

2. Approximately specified information

1. Let us take the case when, instead of the element f , the approximation \tilde{f} of it is specified. Given fixed $\tilde{f} \in N_R$, we put

$$N_\delta(f) = \{\tilde{f} \in F : \|f - \tilde{f}\|_F \leq \delta\}, \quad \delta > 0.$$

We introduce the set of all approximate data:

$$\tilde{N}_R = \bigcup_{\tilde{f}} N_\delta(\tilde{f}), \quad \tilde{f} \in N_R.$$

For the numerical parameters $\delta > 0, R > 0$, we define the set (the element $\tilde{f} \in \tilde{N}_R$ is fixed)

$$U_{\delta, R}(\tilde{f}) = \{u \in M_R : \|Au - \tilde{f}\|_F \leq \delta\}.$$

Obviously, $U_R(f) \subseteq U_{\delta, R}(\tilde{f})$ for any $\delta > 0, R > 0$. We assume that it is known *a priori* that the element u , on which the value of the operator B is evaluated, belongs to the set $U_{\delta, R}(\tilde{f})$.

Let T be any (not necessarily linear) operator, defined on F , and mapping F into V . We characterize the error of approximation of the operator B by the quantity

$$\omega_B(\delta, R, T) = \sup_{\tilde{f}} \sup_u \|Bu - T\tilde{f}\|_V, \quad u \in U_{\delta, R}(\tilde{f}), \quad \tilde{f} \in N_R.$$

We introduce the set

$$U_{\delta, R} = \{u \in M_R : \|Au\|_F \leq \delta\}$$

and the quantity

$$\omega_B(\delta, R) = \sup_u \|Bu\|_V, \quad u \in U_{\delta, R},$$

which we shall assume to be finite for all $\delta > 0, 0 < R < +\infty$.

Theorem 4

Under the above assumptions, for any operator T we have the lower bound

$$\omega_B(\delta, R, T) \geq \omega_B(\delta, R).$$

The proof follows the same lines as the proof of Theorem 1.

Let us emphasize that, in Theorems 1 and 4, all the spaces are assumed to be normed, and the operator B is not necessarily linear.

If $A = E$ and B is linear, Theorems 1 and 4 are the same as the theorems proved in [7]. The following problem was stated in [7]: to find the operator T_{opt} from the condition

$$\omega_B(\delta, R, T_{\text{opt}}) = \inf_T \omega_B(\delta, R, T), \quad (15)$$

when the quasi-optimal (orderwise optimal) operator is defined, i.e., the operator $T_{K \text{ opt}}$ for which the constant $K, 0 < K < +\infty$, exists, such that

$$\omega_B(\delta, R, T_{K \text{ opt}}) \leq K \omega_B(\delta, R, T_{\text{opt}}).$$

It was shown in [7] that the "smoothing" operators, constructed in the following ways, are also quasi-optimal:

a) on the basis of a choice of regularization parameter α from the condition $\|Au_\alpha - \tilde{f}\|_F = \delta$, where u_α is the regularizing family of elements;

b) on the basis of the discrepancy method;

c) on the basis of quasi-solutions;

d) on the basis of a determinate Bayes approach.

The operators corresponding to a) and d) are linear.

If the criterion (15) is chosen, it is difficult to determine the optimal algorithm T_{opt} , since the set $U_{\delta,R}(\tilde{f})$ has a complicated structure. The natural way of the difficulty is to take a different but closely similar criterion, such that the structure of the optimal operator is not influenced by the "geometry" of the set $U_{\delta,R}(\tilde{f})$.

2. Assume that approximate information is given about the element $u \in D$, in the form of an element $\tilde{f} \in \tilde{N}$ where \tilde{N} is some arbitrary set of F . For the numerical parameter $\lambda > 0$ we introduce the functional (see [1])

$$\Phi_\lambda[u, \tilde{f}] = \lambda \|Au - \tilde{f}\|_F^2 + \|Lu\|_G^2.$$

We take the following measures to approximate the operator B :

a) at the point \tilde{f} ,

$$\omega_B(\lambda, T, \tilde{f}) = \sup_u \{ \|Bu - T\tilde{f}\| / \Phi_\lambda^{1/2}[u, \tilde{f}] \}, \quad u \in D;$$

b) in the entire set \tilde{N} ,

$$\omega_B(\lambda, T) = \sup_{\tilde{f}} \omega_B(\lambda, T, \tilde{f}), \quad \tilde{f} \in \tilde{N}.$$

We can state the following problems of optimal approximation of the operator B : to find the operator T_λ , optimal at a point, from the condition

$$\omega_B(\lambda, T_\lambda, \tilde{f}) = \inf_T \omega_B(\lambda, T, \tilde{f}) = \omega_B(\lambda, \tilde{f}); \quad (16)$$

to find the operator P_λ , optimal in \tilde{N} , from the condition

$$\omega_B(\lambda, P_\lambda) = \inf_T \omega_B(\lambda, T) = \omega_B(\lambda, \tilde{N}). \quad (17)$$

Here the infimum is taken over all operators T , defined in the set \tilde{N} .

We shall henceforth assume that H , F , and G are Hilbert spaces. Before proceeding to the solution of problems (16) and (17), consider the following auxiliary regularized problem: to find the element $u_\lambda \in D$ such that

$$\Phi_\lambda[u_\lambda, g] = \inf_{u \in D} \Phi_\lambda[u, g], \quad g \in F. \quad (18)$$

It was shown in [7] that, if the complementarity condition holds and the operators A and L are jointly closed, the element u_λ exists and is unique for any $g \in F$. We can then define, on the basis of a solution of problem (18), a single-parameter family of "smoothing" operators S_λ , such

that the element $u_\lambda : S_\lambda g = u_\lambda \forall g \in F$, is associated with the element g . We shall call the element u_λ the regularized pseudo-solution. For the case $\bar{D} = H$, and a bounded operator A , the operator S_λ can be written explicitly as

$$S_\lambda = \lambda(\lambda A^* A + L^* L)^{-1} A^*.$$

The operator S_λ is obviously linear and bounded in F .

Theorem 5

If the operator B is linear, then, given any $\lambda > 0$, a solution of problem (16) is provided by the linear operator

$$T_\lambda = B S_\lambda.$$

If λ is independent of \tilde{f} , the operator T_λ is also a solution of problem (17). Here,

$$\omega_B(\lambda, \tilde{f}) = \sup_u \{ \|Bu - B\tilde{u}_\lambda\|_V / \Phi_\lambda^{1/2}[u, \tilde{f}] \}, \quad u \in D, \quad (19)$$

$$\omega_B(\lambda, \tilde{N}) = \sup_{\tilde{f}} \omega_B(\lambda, \tilde{f}), \quad \tilde{f} \in \tilde{N}, \quad (20)$$

where $\tilde{u}_\lambda = S_\lambda \tilde{f}$.

Proof. It can easily be shown that, for the regularized pseudo-solution \tilde{u}_λ and any $h \in D$, we have

$$\Phi_\lambda[2\tilde{u}_\lambda - h, \tilde{f}] = \Phi_\lambda[h, \tilde{f}].$$

Further, given any operator T and any $h \in D$, we have

$$\begin{aligned} \omega_B(\lambda, T, \tilde{f}) &\geq \max \{ \|B(2\tilde{u}_\lambda - h) - T\tilde{f}\| / \Phi_\lambda^{1/2}[2\tilde{u}_\lambda - h, \tilde{f}], \\ \|Bh - T\tilde{f}\| / \Phi_\lambda^{1/2}[h, \tilde{f}] \} &= \max \{ \|B(2\tilde{u}_\lambda - h) - T\tilde{f}\| / \Phi_\lambda^{1/2}[h, \tilde{f}], \\ \|Bh - T\tilde{f}\| / \Phi_\lambda^{1/2}[h, \tilde{f}] \}. \end{aligned} \quad (21)$$

Since the operator B is linear, we have

$$2B(\tilde{u}_\lambda - h) = B(2\tilde{u}_\lambda - h) - T\tilde{f} + (T\tilde{f} - Bh),$$

whence, applying the triangle inequality,

$$2\|B\tilde{u}_\lambda - Bh\|_V \leq 2\max\{\|B(2\tilde{u}_\lambda - h) - T\tilde{f}\|_V, \|T\tilde{f} - Bh\|_V\}.$$

Hence we obtain from (21), for any $\tilde{f} \in \tilde{N}$ and any T ,

$$\omega_B(\lambda, T, \tilde{f}) \geq \sup_u \{ \|Bu - BS_\lambda \tilde{f}\| / \Phi_\lambda^{1/2}[u, \tilde{f}] \}, \quad u \in D, \quad (22)$$

and the sign of equality holds for $T = BS_\lambda$. This in fact implies that the operator T_λ is optimal at a point.

Now let λ be independent of the choice of the element \tilde{f} . We then take the supremum in both sides of the inequality (22) with respect to $\tilde{f} \in \tilde{N}$. We get

$$\begin{aligned} &\sup_{\tilde{f} \in \tilde{N}} \sup_{u \in D} \{ \|Bu - T\tilde{f}\| / \Phi_\lambda^{1/2}[u, \tilde{f}] \} \\ &\geq \sup_{\tilde{f} \in \tilde{N}} \sup_{u \in D} \{ \|Bu - BS_\lambda \tilde{f}\| / \Phi_\lambda^{1/2}[u, \tilde{f}] \}. \end{aligned} \quad (23)$$

Since the operator BS_λ is independent of u and \tilde{f} , we can substitute $T = BS_\lambda$ on the left side of (2). The equality is then obtained in (23). Consequently, $P_\lambda = T_\lambda$.

Equations (19) and (20) follow from (22) and (23) respectively. The theorem is proved.

It is easily shown that, when the B -complementarity condition holds, $\omega_B(\lambda, \tilde{N})$ is finite for any $\lambda > 0$.

3. The optimality criteria (16) and (17) have a certain universality. First, whatever the choice of the parameter λ , the method based on regularization proves to be optimal in the sense of the criterion. It is natural to try to find an optimal operator which also ensures stable evaluation of the values of Bu . Such an operator can only be found when supplementary *a priori* information is available. A second feature of the criterion is that supplementary *a priori* information (about the set \tilde{N}) does not influence the structure of the optimal operator, and can only affect the choice of the parameter λ . In this way, the construction of optimal operators for different initial data can be reduced to a suitable choice of the parameter λ .

Notice that, in the case of optimization at a point, it is admissible for the parameter λ to be dependent on the function \tilde{f} . For instance, if the number $\delta > 0$ is given, such that $\|Au - \tilde{f}\|_F \leq \delta$, then a parameter $\lambda = \lambda(\delta, \tilde{f})$ can be chosen from the condition $\|A\tilde{u}_\lambda - \tilde{f}\|_F = \delta$. The resulting operator $T_{\lambda(\delta, \tilde{f})}$ will be optimal in the sense of the criterion (16).

In the case of optimization in the set \tilde{N} , the parameter λ has to be independent of the choice of \tilde{f} , and has to be defined for the entire set \tilde{N} . Let $\tilde{N} \equiv \tilde{N}_R$, i.e., the characteristics of \tilde{N} are the numbers $\delta > 0$ and $R > 0$. An example of a choice of λ dependent on the entire set \tilde{N} is $\lambda = \bar{\lambda} = R^2/\delta^2$ (which leads to the so-called determinate Bayes method of regularization [7]). The operator $T_{\bar{\lambda}}$ is optimal in \tilde{N}_R in the sense of the criterion (17).

As we remarked earlier, it was shown in [7] that the operators $T_{\bar{\lambda}}$ and $T_{\lambda(\delta, \tilde{f})}$ are quasi-optimal with respect to the criterion (15), in fact,

$$\sup_{\tilde{f} \in \tilde{N}_R} \sup_{u \in U_{\delta, R}(\tilde{f})} \|Bu - BS_{\lambda}f\|_V \leq 2^{1/2} \omega_B(\delta, R), \quad \lambda = \bar{\lambda}, \lambda(\delta, \tilde{f}).$$

In [7] the conditions under which $\omega_B(\delta, R) \rightarrow 0$ as $\delta \rightarrow 0$ were stated. Obviously, under these conditions the optimal operators $T_{\lambda(\delta, \tilde{f})}$ and $T_{\bar{\lambda}}$ ensure stable computation of the values of the operator B .

Some of the present results were published in [13].

Translated by D. E. Brown.

REFERENCES

1. TIKHONOV, A. N., On the solution of ill posed problems and a method of regularization, *Dokl. Akad. Nauk SSSR*, **151**, No. 3, 501-504, 1963.
2. STECHKIN, S. B., Best approximation of linear operators, *Matem. zametki*, **1**, No. 2, 137-148, 1967.
3. BAKHVALOV, N. S., On the optimality of linear methods of approximating operators in convex function classes, *Zh. vychisl. Mat. mat. Fiz.*, **11**, No. 4, 1014-1018, 1971.
4. MARCHUK, A. G., and OSIPENKO, K. YU., Best approximation of functions specified with an error at a finite number of points, *Matem. zametki*, No. 3, 359-368, 1975.
5. REINSCH, C., Two extensions of the Sard-Schoenberg theory of best approximations, *SIAM J. Numer. Analysis*, **11**, No. 1, 45-51, 1974.

6. MUNTEANU, M. J., Generalized smoothing spline functions for operators, *SIAM J. Numer. Analysis*, **11**, No. 1, 28–34, 1973.
7. MOROZOV, V. A., *Regular methods for solving ill posed problems* (Regulyarnye metody resheniya nekorrektno postavlennykh zadach), Izd-vo MGU, Moscow, 1974.
8. STRAKHOV, V. N., On the solution of linear ill-posed problems in Hilbert space, *Differents. ur-niya*, **6**, No. 8, 1490–1495, 1970.
9. IVANOV, V. K., and KOROLYUK, T. I., On estimation of the error when solving linear ill posed problems, *Zh. vychisl. Mat. mat. Fiz.*, **9**, No. 1, 30–41, 1969.
10. IVANOV, V. V., On accuracy-optimal algorithms for approximate solution of operator equations of the 1st kind, *Zh. vychisl. Mat. mat. Fiz.*, **15**, No. 1, 3–11, 1975.
11. MOROZOV V. A., On a stable method of computing the values of an unbounded operator, *Dokl. Akad. Nauk SSSR*, **185**, No. 2, 267–270, 1969.
12. MOROZOV, V. A., Convergence of an approximate method for solving operator equations of the 1st kind, *Zh. vychisl. Mat. mat. Fiz.*, **13**, No. 1, 3–17, 1973.
13. MOROZOV, V. A., and GREBENNIKOV, A. I., On the optimal approximation of operators, *Dokl. Akad. Nauk SSSR*, **223**, No. 6, 1307–1310, 1975.

ON AN ITERATIVE PROJECTION ALGORITHM FOR SOLVING ILL POSED PROBLEMS WITH AN APPROXIMATELY SPECIFIED OPERATOR*

V. P. TANANA

Sverdlovsk

(Received 11 February 1975; revised 21 April 1975)

A PROJECTION algorithm of the iterative type is proposed, for solving approximately linear operator equations of the 1st kind with an approximately specified right-hand side and an operator in Hilbert space.

An original method for solving an operator equation of the 1st kind with a disturbed operator, representing an extension of the discrepancy method [3–5], was described in [1], in the context of compact embedding [2]. A similar though somewhat different approach to the solution of ill posed problems with a disturbed operator was considered in [6].

The basic idea of the method described in [1] lies in reducing the problem of the approximate solution of the operator equation to a variational problem with non-linear (non-convex) constraints; but the solution of this latter problem is quite difficult and requires the development of special methods.

In the present paper the method described in [1] is justified for linear operator equations without compact embedding; this is extremely important when solving the converse problem of gamma logging of wells [7], in which the exact solution (radioactive element content) is often a discontinuous function, about which no *a priori* information is available. Further, it is shown, under the same assumptions, that the method described in [1] can be reduced to a method of Tikhonov regularization [8], with a parameter α chosen according to a generalized discrepancy principle [9, 10]; and finally, an iterative projection algorithm is outlined and proved for realizing this method.

*Zh. vychisl. Mat. mat. Fiz., **17**, 1, 15–23, 1977.

1. Statement of the problem and method of solution

Let X be an E space (see [11]), Y a Banach space, and A a linear one-to-one continuous operator mapping X into Y . We consider the operator equation of the 1st kind

$$Ax=y, \quad x \in X, \quad y \in Y. \quad (1.1)$$

Assume that, for $y = y_0$, the equation has a solution x_0 , but that y_0 and A are unknown to us; we only know the quantity y_δ such that $\|y_\delta - y_0\| < \delta$, and the linear continuous operator A_h , mapping X into Y and satisfying the condition $\|A_h - A\| \leq h$, where δ and h are positive numerical parameters, and $\|y_\delta\| > \delta + \|x_0\|h$, $h < \|A_h\|$. Knowing y_δ and A_h , we want to construct the approximate solution $x_{\delta h}$ of Eq. (1.1), satisfying the condition $x_{\delta h} \rightarrow x_0$ as $\delta + h \rightarrow 0$.

The method of solution amounts to reducing the problem of the approximate solution of Eq. (1.1) to the variational problem

$$\inf \{ \|x\|^{\gamma_1} \mid \|A_h x - y_\delta\| \leq \delta + \|x\|h \}, \quad \gamma_1 > 1, \quad (1.2)$$

see [1].

Theorem 1

Let the domain of values of the operator A be everywhere dense in Y , $R_{A_h} = Y$. Then the variational problem (1.2) is equivalent to the problem

$$\inf_{x \in X} \{ \|x\|^{\tau_1} \mid \|A_h x - y_\delta\| \leq \delta + \tau h \} \quad (1.3)$$

with the connection

$$\|x_{\delta h}^{\tau_1}\| = \tau. \quad (1.4)$$

Proof. Let $0 \leq \tau \leq (\|y_\delta\| - \delta)/h$. Consider the function $\varphi(\tau) = \tau - \|x_{\delta h}^{\tau_1}\|$, where $x_{\delta h}^{\tau_1}$ is the solution of the problem (1.3). The function $\varphi(\tau)$ is obviously monotonically increasing, and satisfies the end conditions $\varphi(0) < 0$ and $\varphi((\|y_\delta\| - \delta)/h) > 0$. Hence a unique τ_0 exists, at which $\varphi(\tau_0) = 0$. Problem (1.3), (1.4) is thus uniquely solvable. We denote the solution of the problem (1.3), (1.4) by $x_{\delta h}^{\tau_0}$ and we aim to show that the element $x_{\delta h}^{\tau_0}$ is the unique solution of problem (1.2). For this, we observe that $x_{\delta h}^{\tau_0}$ satisfies the constraints in the problem (1.2);

Assume that $x_{\delta h}^{\tau_0}$ is not a solution of problem (1.2). There will then be an element $\bar{x} \in X$ such that $\|A_h \bar{x} - y_\delta\| \leq \delta + \|\bar{x}\|h$ and $\|\bar{x}\| < \|x_{\delta h}^{\tau_0}\|$, where $d > 0$. The element \bar{x} will then satisfy the inequality $\|A_h \bar{x} - y_\delta\| \leq \delta + \|x_{\delta h}^{\tau_0}\|h$, and noting that

$$\|x_{\delta h}^{\tau_0}\| = \inf_{x \in X} \{ \|x\| \mid \|A_h x - y_\delta\| \leq \delta + \|x_{\delta h}^{\tau_0}\|h \},$$

we obtain $\|\bar{x}\| \geq \|x_{\delta h}^{\tau_0}\|$, but this contradicts the inequality $\|\bar{x}\| < \|x_{\delta h}^{\tau_0}\|$.

To prove the uniqueness, assume that \tilde{x} is another solution of problem (1.2). This solution will satisfy the conditions $\|\tilde{x}\| = \|x_{\delta h}^{\tau_0}\|$ and $\|A_h \tilde{x} - y_\delta\| \leq \delta + \|x_{\delta h}^{\tau_0}\|h$, but X is strictly convex, so that, on the basis of [12], we have $\tilde{x} = x_{\delta h}^{\tau_0}$. This proves the theorem.

It follows from Theorem 1 that, under our assumptions, problem (1.2) is uniquely solvable. Recalling the results of [13], we can conclude from Theorem 1 that the variational problem (1.2) is equivalent to the problem

$$\inf_{x \in X} \{ \|A_h x - y_\delta\|^{\gamma_2} + \alpha \|x\|^{\gamma_1} \}$$

with the connection

$$\|A_h x_{\delta h} - y_\delta\| = \delta + \|x_{\delta h}\| h, \quad \gamma_2 \geq 1.$$

Henceforth, the approximate solution of Eq. (1.1) (the solution of the variational problem (1.2)) will be denoted by $x_{\delta h}$.

Theorem 2

The element $x_{\delta h}$ is convergent to x_0 as $\delta + h \rightarrow 0$.

Proof. Assume the contrary, i.e., a sequence $x_{\delta_k h_k}$ exists, such that $\delta_k + h_k \rightarrow 0$ as $k \rightarrow \infty$ and

$$\|x_{\delta_k h_k} - x_0\| \geq d > 0. \quad (1.5)$$

Since $\|x_{\delta_k h_k}\| \leq \|x_0\|$ for any k , and the space X is reflexive, the sequence $\{x_{\delta_k h_k}\}$ must be weakly compact. It can therefore be assumed without loss of generality that $x_{\delta_k h_k} \xrightarrow{w} \hat{x}$ as $k \rightarrow \infty$.

On the other hand, $\|A x_{\delta_k h_k} - y_{\delta_k}\| \leq 2(\delta_k + \|x_0\| h_k)$. Hence $A x_{\delta_k h_k} \rightarrow y_0$ as $k \rightarrow \infty$.

Recalling that the operator A is linear and continuous, we get $x = x_0$. In view of the fact that $\|x_{\delta_k h_k}\| \leq \|x_0\|$ for any k , and the fact that X is an E space, we have $x_{\delta_k h_k} \rightarrow x_0$ as $k \rightarrow \infty$; but this contradicts (1.5).

2. Method of finite-dimensional approximations of the approximate solution $x_{\delta h}$

We consider the increasing chain of finite-dimensional subspaces of the space X

$$X_1 \subset \dots \subset X_n \subset \dots \subset X$$

such that

$$\overline{\bigcup_{n=1}^{\infty} X_n} = X, \quad (2.1)$$

and the variational problem

$$\inf \{ \|x\|^{\gamma_1} \mid x \in X_n, \|A_h x - y_\delta\| \leq \delta + \|x\| h \}. \quad (2.2)$$

Theorem 3

Problem (2.2) is uniquely solvable for sufficiently large n .

Proof. Let us first show that the set $\{x \in X_n \mid \|A_h x - y_\delta\| \leq \delta + \|x\| h\}$ is not empty. For this, we note that, in view of condition (2.1), there exist X_N and $x' \in X_N$ such that $\|x' - x_0\| < \delta - \|y_\delta - y_0\|$,

but then, $\|A_h x' - y_\delta\| \leq \delta + \|x'\|h$ and hence, for all $n \geq N$, the sets $\{x \in X_n \mid \|A_h x - y_\delta\| \leq \delta + \|x\|h\}$ are not empty; and by Theorem 1, for $n \geq N$ problem (2.2) is uniquely solvable. This proves the theorem.

Henceforth the solution of problem (2.2) will be denoted by $x_{\delta h}^n$.

Theorem 4

The solutions $x_{\delta h}^n$ converge to $x_{\delta h}$ as $n \rightarrow \infty$.

Proof. Assume that the convergence $x_{\delta h}^n \rightarrow x_{\delta h}$, does not hold, i.e., a sequence $x_{\delta h}^{n_k}$ exists such that

$$\|x_{\delta h}^{n_k} - x_{\delta h}\| \geq d > 0. \quad (2.3)$$

Since the sequence $x_{\delta h}^{n_k}$ is bounded, it is weakly compact, and we can thus assume without loss of generality that

$$x_{\delta h}^{n_k} \xrightarrow{w} \hat{x} \text{ as } k \rightarrow \infty.$$

Since the operator A_h is linear, given any $\epsilon > 0$ we can find an element $\bar{x}_{\delta h}$ such that $\|\bar{x}_{\delta h}\| < \|x_{\delta h}\| + \epsilon$ and $\|A_h \bar{x}_{\delta h} - y_\delta\| < \delta + \|x_{\delta h}\|h$.

Recalling the continuity of the operator A_h and the property of the system of subspaces X_n , we can assert the existence of a sequence $\{x_{n_k}\}$, $x_{n_k} \in X_{n_k}$, such that $\|x_{n_k}\| = \|\bar{x}_{\delta h}\|$, $x_{n_k} \rightarrow \bar{x}_{\delta h}$ as $n_k \rightarrow \infty$ and $\|A_h x_{n_k} - y_\delta\| < \delta + \|x_{\delta h}\|h$. Then, $\|A_h x_{n_k} - y_\delta\| < \delta + \|x_{n_k}\|h$, and hence $\|x_{\delta h}^{n_k}\| \leq \|x_{n_k}\|$. Consequently, $\lim_{n_k \rightarrow \infty} \|x_{\delta h}^{n_k}\| \leq \|x_{\delta h}\| + \epsilon$, and since ϵ is arbitrary, we have

$$\lim_{n_k \rightarrow \infty} \|x_{\delta h}^{n_k}\| \leq \|x_{\delta h}\|. \quad (2.4)$$

Since $\|A_h x_{\delta h}^{n_k} - y_\delta\| \leq \delta + \|x_{\delta h}^{n_k}\|h$ for any k , and the operator A_h is linear and continuous, we have

$$\|A_h \hat{x} - y_\delta\| \leq \delta + \|x_{\delta h}^{n_k}\|h \quad \forall n_k. \quad (2.5)$$

From (2.4) and (2.5) we obtain

$$\|A_h \hat{x} - y_\delta\| \leq \delta + \|x_{\delta h}\|h, \quad (2.6)$$

and recalling that $\|x_{\delta h}\| = \inf_{x \in X} \{\|x\| \mid \|A_h x - y_\delta\| \leq \delta + \|x\|h\}$, we find from (2.4) and (2.6) that $\hat{x} = x_{\delta h}$. Recalling that X is an E space, we find from this last equation and (2.4) that

$$x_{\delta h}^{n_k} \rightarrow x_{\delta h} \text{ as } n_k \rightarrow \infty,$$

which contradicts (2.3). This proves the theorem.

3. An iterative method of solving problem (2.2) for $\gamma_1 = 2$

Let $X = Y = H$, where H is a separable Hilbert space, and let A_h be a linear one-to-one continuous operator, mapping X into Y , with domain of values R_{A_h} , everywhere dense in Y ; the subspace X_n satisfies the condition $\rho(y_\delta, A_h X_n) < \delta$.

Let x_k be the k -th iteration, which satisfies the conditions $x_k \in X_n$ and $\|A_h x_k - y_\delta\| = \delta + \|x_k\|h$.

Then we obtain the $(k+1)$ -th iteration x_{k+1} by solving the problem

$$\inf \{ \|x\|^2 \mid x \in L(x_k, A_{h_n}'(A_h x_k - y_\delta)); \|A_h x - y_\delta\| = \delta + \|x\|h \}. \quad (3.1)$$

Here, $L(\tilde{x}_k, A_{h_n}'(A_h x_k - y_\delta))$ is the linear hull, stretched over the elements x_k and $A_{h_n}'(A_h x_k - y_\delta)$, and A_{h_n}' is the operator adjoint to the operator A_{h_n} , which is the contraction of the operator A_h from the space X onto X_n .

Lemma 1

If $\rho(y_\delta, A_h X_n) < \delta$ and $\|A_h \bar{x} - y_\delta\| = \|x\|h$, then $A_{h_n}'(A_h \bar{x} - y_\delta) \neq 0$.

Proof. Under the conditions of the lemma,

$$A_h \bar{x} - y_\delta = (A_h \bar{x} - \text{pr}(A_h X_n, y_\delta)) + (\text{pr}(A_h X_n, y_\delta) - y_\delta), \quad (3.2)$$

where $\text{pr}(A_h X_n, y_\delta)$ is the metric projection of the element y_δ onto the subspace $A_h X_n$ and $A_h \bar{x} - \text{pr}(A_h X_n, y_\delta) \neq 0$.

By definition of the adjoint operator, for any $x \in X_n$ we have

$$(A_{h_n}'(A_h \bar{x} - y_\delta), x) = (A_h \bar{x} - y_\delta, A_h x).$$

Hence, recalling the decomposition (3.2), we obtain

$$(A_{h_n}'(A_h \bar{x} - y_\delta), x) = (A_h \bar{x} - \text{pr}(A_h X_n, y_\delta), A_h x).$$

From this equation we have

$$(A_{h_n}'(A_h \bar{x} - y_\delta), \tilde{x}) \neq 0,$$

where $\tilde{x} = \bar{x} - A_h^{-1}(\text{pr}(A_h X_n, y_\delta))$. Hence $A_{h_n}'(A_h \bar{x} - y_\delta) \neq 0$; this proves the lemma.

Since the hyperplane $\{y \mid (A_h \bar{x} - y_\delta, y) = (A_h \bar{x} - y_\delta, A_h \bar{x})\}$ supports the sphere $S_{\delta + \|x\|h}$ ($y_\delta = \{y \mid \|y - y_\delta\| \leq \delta + \|\bar{x}\|h\}$) at the point $A_h \bar{x}$, the corresponding hyperplane

$$\bar{G} = \{x \in X_n \mid (A_{h_n}'(A_h \bar{x} - y_\delta), x) = (A_{h_n}'(A_h \bar{x} - y_\delta), \bar{x})\}$$

will support the set $\Omega_{\delta h} = \{x \in X_n \mid \|A_h x - y_\delta\| \leq \delta + \|\bar{x}\|h\}$ at the point \bar{x} .

Let x_k be a point satisfying the condition $x_k \in X_n, \|A_h x_k - y_\delta\| = \delta + \|x_k\|h$; then the hyperplane supporting the set $\{x \in X_n \mid \|A_h x - y_\delta\| \leq \delta + \|x_{k+1}\|h\}$ at the $(k+1)$ -th iteration x_{k+1} (see (3.1)) will either separate the set $\{x \in L(x_k, A_{h_n}'(A_h x_k - y_\delta)) \mid \|A_h x - y_\delta\| \leq \delta + \|x_n\|h\}$ and the point 0, or will contain the point 0.

Lemma 2

If x_k is the k -th iteration and $x_k \neq x_{\delta h}^n$, while x_{k+1} is the $(k+1)$ -th iteration, see (3.1), then $\|x_{k+1}\| < \|x_k\|$:

Proof. Recalling the remark made above, it can be assumed without loss of generality that $A_{h_n}'(A_h x_h - y_\delta) \neq \lambda x_h$ for any values of λ , since otherwise $x_{h+1} = x_h$, and hence $\|x_h\|^2 = \inf \{\|x\|^2 \mid x \in X_n, \|A_h x - y_\delta\| \leq \delta + \|x_h\|h\}$, which contradicts the hypothesis of the Lemma 2.

Obviously, the projection $\text{pr}(G_k, 0)$ of the point 0 onto the hyperplane G_k will satisfy the conditions:

$$\text{pr}(G_k, 0) \in \inf \{\|x\| \mid x \in X_n, \|A_h x - y_\delta\| \leq \delta + \|x_h\|h\},$$

$$\text{pr}(G_k, 0) = \lambda_0 A_{h_n}'(A_h x_h - y_\delta),$$

where λ_0 is a number, and G_k is the hyperplane supporting the set $\{x \in X_n \mid \|A_h x - y_\delta\| \leq \delta + \|x_h\|h\}$ at the point x_k .

We choose a number $\epsilon \neq 0$ such that the hyperplane G_k does not separate the point $\text{pr}(G_k, 0) + \epsilon A_{h_n}'(A_h x_h - y_\delta)$ and the set $\{x \in X_n \mid \|A_h x - y_\delta\| \leq \|A_h x_h - y_\delta\|\}$ and such that $\|\text{pr}(G_k, 0) + \epsilon A_{h_n}'(A_h x_h - y_\delta)\| < \|x_h\|$, and we consider the interval $T = \{\alpha x_k + (1-\alpha)[\text{pr}(G_k, 0) + \epsilon A_{h_n}'(A_h x_h - y_\delta)], 0 \leq \alpha \leq 1\}$, containing the points $(G_k, 0) + \epsilon A_{h_n}'(A_h x_h - y_\delta)$ and x_k .

Then, since the boundary of the set $\{x \in X_n \mid \|A_h x - y_\delta\| \leq \|A_h x_h - y_\delta\|\}$ is smooth, there will be a point $\tilde{x}_0 = \alpha_0 x_k + (1-\alpha_0)[\text{pr}(G_k, 0) + \epsilon A_{h_n}'(A_h x_h - y_\delta)]$, $0 < \alpha_0 < 1$, such that $\|A_h \tilde{x}_0 - y_\delta\| < \|A_h x_h - y_\delta\|$. Hence there exists a point $\tilde{x}_0 \in L(x_k, A_{h_n}'(A_h x_h - y_\delta))$, $\|\tilde{x}_0\| = \|x_k\|$, and $\|A_h \tilde{x}_0 - y_\delta\| < \|A_h x_h - y_\delta\|$, where $\|A_h x_h - y_\delta\| = \delta + \|x_h\|h$. But then, we can choose a number μ_0 , $|\mu_0| < 1$ such that $\|A_h(\mu_0 \tilde{x}_0) - y_\delta\| \leq \delta + \|\mu_0 \tilde{x}_0\|h$, whence it follows that $\|x_{h+1}\| < \|x_k\|$. This proves the lemma.

Lemma 3

Let the sequences $\{x_k\}$, $\{x_k'\} \subset X_n$ be such that $L(x_k, x_k') \cap \Omega_{\delta h_k}^n \neq \emptyset$, where $\Omega_{\delta h_k}^n = \{x \in X_n \mid \|A_h x - y_\delta\| \leq \delta + a_k h\}$, $a_{k+1} \leq a_k$, $a_k \rightarrow a_0$ as $k \rightarrow \infty$, and $x_k \rightarrow x$, $x_k' \rightarrow x'$ as $k \rightarrow \infty$, while $x \neq \lambda x'$ for any values of λ . Then,

$$\lim_{k \rightarrow \infty} \|\bar{x}_k\| \leq \|\bar{x}\|,$$

where $\|\bar{x}_k\|^2 = \inf \{\|x\|^2 \mid x \in L(x_k, x_k') \cap \Omega_{\delta h_k}^n\}$, $\|\bar{x}\|^2 = \inf \{\|x\|^2 \mid x \in L(x, x') \cap \Omega_{\delta h_0}^n\}$.

Proof. Assume the contrary, i.e.,

$$\lim_{k \rightarrow \infty} \|\bar{x}_k\| > \|\bar{x}\|.$$

Then, a subsequence $\{\bar{x}_{k_l}\}$ exists, such that

$$\|\bar{x}_{k_l}\| > \|\bar{x}\| + d, \quad (3.3)$$

where d is a positive number. Assume that $0 \in G(\bar{x})$, where $G(\bar{x}) \subset X_n$ is the hyperplane supporting the set $\Omega_{\delta h_0}^n$ at the point \bar{x} , which, since the boundary of the set $\Omega_{\delta h_0}^n$ is smooth, is the tangent hyperplane to the set $\Omega_{\delta h_0}^n$ at the point \bar{x} , while in view of the fact that $0 \in G(\bar{x})$, i.e., the intersection $L(x, x') \cap \Omega_{\delta h_k}^n$ consists of more than one point, we have $L(x_k, x_k') \cap \Omega_{\delta h_0}^n \neq \emptyset$ for sufficiently large k . Denote by \bar{x}_k^0 the point satisfying the relation

$$\|\bar{x}_k^0\|^2 = \inf \{\|x\|^2 \mid x \in L(x_k, x_k') \cap \Omega_{\delta h_0}^n\}.$$

Since $\Omega_{\delta h_k}^n \supset \Omega_{\delta h_0}^n$, we have $\|\bar{x}_k\| < \|\bar{x}_k^0\|$.

Consider the metric projection \tilde{x}_{k_l} of the element \bar{x} onto the set $L(x_{k_l}, x'_{k_l})$. We can assume without loss of generality that $\tilde{x}_{k_l} \rightarrow \bar{x}$ as $k_l \rightarrow \infty$, where $\tilde{x} \in L(x, x')$.

We then consider the sequence of points $\{z_{k_l}\} \subset G(\bar{x})$, where $z_{k_l} = \mu_{k_l} \tilde{x}_{k_l}$, μ_{k_l} is a number. Obviously,

$$\|z_{k_l} - \bar{x}\| = \|\tilde{x}_{k_l} - \bar{x}\| / \cos \alpha_{k_l}, \quad (3.4)$$

where α_{k_l} is the angle between the vectors $\tilde{x}_{k_l} - \bar{x}$ and $z_{k_l} - \bar{x}$,

$$\cos \alpha_{k_l} = \frac{(\tilde{x}_{k_l} - \bar{x}, z_{k_l} - \bar{x})}{\|\tilde{x}_{k_l} - \bar{x}\| \|z_{k_l} - \bar{x}\|}.$$

Since $0 \in G(\bar{x})$, we have $\sup \{(-\bar{x}, x - \bar{x}) \mid x \in G(\bar{x}), \|x - \bar{x}\| = 1\} < \|x\|$, and hence the angle $\angle(G(\bar{x}), \bar{x})$ between the hyperplane $G(\bar{x})$ and the element \bar{x} is positive. We can assume without loss of generality that $\alpha_{k_l} \rightarrow \alpha_0$ and $|\alpha_0| < \pi/2 - \angle(G(\bar{x}), \bar{x})$. Hence, for sufficiently large k_l , we have $|\alpha_{k_l}| < \pi/2 - \alpha_1$, where $\alpha_1 > 0$.

From this and (3.4) we have

$$z_{k_l} \rightarrow \bar{x} \quad \text{as} \quad k_l \rightarrow \infty. \quad (3.5)$$

We consider the sequence $\{\tilde{z}_{k_l}\}$, where $\tilde{z}_{k_l} = \nu_{k_l} \tilde{x}_{k_l}$, ν_{k_l} is a number and $\|A_h \tilde{z}_{k_l} - y_\delta\| = \|A_h \bar{x} - y_\delta\|$. Since the hyperplane $G(\bar{x})$ is tangential to the set $\Omega_{\delta h_0}^n$ at the point \bar{x} , we have

$$\|\tilde{z}_{k_l} - z_{k_l}\| \rightarrow 0 \quad \text{as} \quad k_l \rightarrow \infty.$$

It follows from this and (3.4) that

$$\|\tilde{z}_{k_l}\| \rightarrow \|\bar{x}\| \quad \text{as} \quad k_l \rightarrow \infty. \quad (3.6)$$

Since $\tilde{z}_{k_l} \in L(x_{k_l}, x'_{k_l}) \cap \Omega_{\delta h_0}^n$, a $\|x_{k_l}^0\|^2 = \inf \{\|x\|^2 \mid x \in L(x_{k_l}, x'_{k_l}) \cap \Omega_{\delta h_0}^n\}$, then $\|x_{k_l}^0\| \leq \|\tilde{z}_{k_l}\|$, and hence $\|\bar{x}_{k_l}\| \leq \|\tilde{z}_{k_l}\|$, and in the light of (3.6) for sufficiently large k , we have $\|\bar{x}_{k_l}\| \leq \|\bar{x}\| + \alpha/2$, which contradicts (3.3).

Assume that $0 \in G(x)$, then,

$$L(x, x') \cap \Omega_{\delta h_0}^n = \{\bar{x}\},$$

since otherwise, $\|x\|^2 \neq \inf \{\|x\|^2 \mid x \in L(x, x') \cap \Omega_{\delta h_0}^n\}$. Noting that $x_{k_l} \rightarrow x$, $x'_{k_l} \rightarrow x'$, and that the operator A_h is continuous, we obtain $A_h x_{k_l} \rightarrow A_h x$ and $A_h x'_{k_l} \rightarrow A_h x'$. Since $a_{k_l} \rightarrow a_0$ as $k_l \rightarrow \infty$, the sequence of sets $S_{\delta+a_{k_l}h}^n(y_\delta)$ is β -convergent to the set $S_{\delta+a_0h}^n(y_\delta)$ (see [14]), and hence the sequence of sets $L(A_h x_{k_l}, A_h x'_{k_l}) \cap S_{\delta+a_{k_l}h}^n(y_\delta)$ is β -convergent to the set $L(A_h x, A_h x') \cap S_{\delta+a_0h}^n(y_\delta)$, i.e.,

$$\sup \{\|y - A_h \bar{x}\| \mid y \in L(A_h x_{k_l}, A_h x'_{k_l}) \cap S_{\delta+a_{k_l}h}^n(y_\delta)\} \rightarrow 0 \quad (3.7)$$

as $k_l \rightarrow \infty$, where $S_{\delta+a_{k_l}h}^n(y_\delta) = \{y \in A_h X_h : \|y - y_\delta\| \leq \delta + a_{k_l}h\}$. Since the operator A_h^{-1} is continuous on $A_h X_h$, the sequence of sets $A_h^{-1}(L(A_h x_{k_l}, A_h x'_{k_l}) \cap S_{\delta+a_{k_l}h}^n(y_\delta))$ will, in the light of (3.6) and (3.7), be α -convergent to the element \bar{x} (see [14]). But then, on the basis of the results of [14], and the fact that

$$\|\bar{x}_{k_l}\|^2 = \inf \{\|x\|^2 \mid x \in A_h^{-1}(L(A_h x_{k_l}, A_h x'_{k_l}) \cap S_{\delta+a_{k_l}h}^n(y_\delta))\},$$

we obtain $\bar{x}_{k_l} \rightarrow \bar{x}$ as $k_l \rightarrow \infty$, which contradicts (3.3). This proves the lemma.

Theorem 5

The successive iterations x_k (see (3.1)) are convergent to $x_{\delta h}^n$ as $k \rightarrow \infty$.

Proof. Since $\|x_{k+1}\| \leq \|x_k\|$, we have

$$\|x_k\| \rightarrow a \text{ as } k \rightarrow \infty \text{ and } \|x_k\| \geq a, \quad (3.8)$$

where a is a number.

Assume that $a > \|x_{\delta h}^n\|$. Since the sequence $\{x_k\}$ is bounded and belongs to the finite-dimensional space X_n , it is compact, and hence we can extract from it a convergent subsequence. Let

$$x_{k_l} \rightarrow \bar{x} \text{ as } k_l \rightarrow \infty. \quad (3.9)$$

Then, $\|\bar{x}\| = a$. Since the hyperplane G_{k_l} , supporting the set $\Omega_{\delta h k}^n$, at the point x_{k_l} , either separates the set $\{x \in X_n \mid \|A_h x - y_\delta\| \leq \delta + \|x_{k_l}\| h\}$ and the point 0, or contains the point 0, then the hyperplane supporting the set $\Omega_{\delta h \|\bar{x}\|}^n = \{x \in X_n \mid \|A_h x - y_\delta\| \leq \delta + \|\bar{x}\| h\}$ at the point \bar{x} will also satisfy this condition.

Hence $A_{h n}'(A_h \bar{x} - y_\delta) \neq \lambda \bar{x}$ for any value of λ . Hence, by Lemma 3,

$$\|\tilde{x}\|^2 < \|\bar{x}\|^2, \quad (3.10)$$

where $\|\tilde{x}\|^2 = \inf \{\|x\|^2 \mid x \in L(\bar{x}, A_{h n}'(A_h \bar{x} - y_\delta)) \cap \Omega_{\delta h \|\bar{x}\|}^n\}$. Since the operator $A_{h n}'$ is continuous, we have

$$A_{h n}'(A_h x_{k_l} - y_\delta) \rightarrow A_{h n}'(A_h \bar{x} - y_\delta) \text{ as } k_l \rightarrow \infty, \quad (3.11)$$

while on the basis of Lemma 1 we have $A_{h n}'(A_h \bar{x} - y_\delta) \neq 0$. But then, by Lemma 3, it follows from (3.9) and (3.11) that $\lim_{k_l \rightarrow \infty} \|x_{k_l}\| \leq \|\tilde{x}\|$, where $\|\tilde{x}\|^2 = \inf \{\|x\|^2 \mid x \in L(x_{k_l}, A_{h n}'(A_h x_{k_l} - y_\delta)) \cap \Omega_{\delta h k_l}^n\}$. If $x_{k_l} = x_{k_l+1}$ and (3.10) is satisfied, we have $\|x_k\| < a$ for sufficiently large k , which contradicts (3.8). Hence $\|x_k\| \rightarrow \|x_{\delta h}^n\|$ as $k \rightarrow \infty$, and hence $x_k \rightarrow x_{\delta h}^n$. This proves the theorem.

Translated by D. E. Brown

REFERENCES

1. GONCHARSKII, A. V., LEONOV, A. S., and YAGOLA, A. G., On a regularizing algorithm for ill posed problems with approximately specified operator, *Zh. vychisl. Mat. mat. Fiz.*, **12**, No. 6, 1592-1594, 1972.
2. TIKHONOV, A. N., On nonlinear equations of the first kind, *Dokl. Akad. Nauk SSSR*, **161**, No. 5, 1023-1026, 1965.
3. PHILLIPS, D. L., A technique for the numerical solution of integral equations of the first kind, *J. Assoc. Comput. Machinery*, **9**, No. 1, 84-97, 1962.
4. DOMBROVSKAYA, I. N., and IVANOV, V. K., On theory of some linear equations in abstract spaces, *Sibirskii matem. zh.*, **6**, No. 3, 499-508, 1965.

5. IVANOV, V. K., On the approximate solution of operator equations of the first kind, *Zh. vychisl. Mat. mat. Fiz.*, **6**, 1089–1094, 1966.
6. MOROZOV, V. A., On a new approach to the solution of linear equations of the 1st kind with an approximate operator, *Proc. of the First Conference of Young Scientists of the Computational Mathematics Dept.*, Izd-vo Mosk. un-ta, 22–28, 1973.
7. DAVYDOV, YU. B., Estimation of the stability of the solution of a linear operator equation of the first kind, obtained by the approximate operator method, *Matem. zap. Ural'skii un-t*, No. 4, 19–26, 1970.
8. TIKHONOV, A. N., On the solution of ill posed problems and a method of regularization, *Dokl. Akad. Nauk SSSR*, **151**, No. 3, 501–504, 1963.
9. GONCHARSKII, A. V., LEONOV, A. S., and YAGOLA, A. G., A generalized discrepancy principle, *Zh. vychisl. Mat. mat. Fiz.*, **13**, No. 2, 294–302, 1973.
10. GONCHARSKII, A. V., LEONOV, A. S., and YAGOLA, A. G., On regularization of ill posed problems with an approximately specified operator, *Zh. vychisl. Mat. mat. Fiz.*, **14**, No. 4, 1022–1027, 1974.
11. KY FAN and GLICKSBERG, J., Some geometric properties of spheres in a normed linear space, *Duke Math. J.*, **25**, No. 4, 553–568, 1958.
12. KLEE, V. L., Convexity of Chebyshev sets, *Math. Ann.*, **142**, No. 3, 292–304, 1961.
13. VASIN, V. V., Ill posed problems in B spaces and their approximate solution by a variational method, Diss. kand. fiz.-matem. n., Sverdlovsk, In-t Matem. i Mekhan., 1970.
14. LISKOVETS, O. A., Ill posed problems and the stability of quasi-solutions, *Sibirskii matem. zh.*, **10**, No. 2, 273–385, 1969.

A NEW TRUNCATION PROCEDURE IN THE BAZLEY–FOX METHOD*

L. T. POZNYAK

Leningrad

(Received 4 May 1975)

WHEN the Bazley–Fox method is used in the standard form, complicated transcendental equations have to be solved in order to obtain a lower bound for the eigenvalues of a self-adjoint positive definite operator A with a discrete spectrum. While, to avoid this difficulty, Bazley and Fox supplemented their method with several devices, it turned out that the devices do not provide good convergence in certain important classes of problem. The present paper offers a new means of simplifying the approximate equations, to which the Bazley–Fox method leads. It reduces finding lower bounds for the eigenvalues of the operator A to a problem of linear algebra, and has good velocity characteristics.

1. Introduction

Suppose we are given in separable Hilbert space H with scalar product (\cdot, \cdot) and norm $|\cdot|$, the self-adjoint positive definite operator (pdo) A with discrete spectrum; we pose the problem of finding the eigenvalue $\{\lambda_i\}$ of A . As usual, we assume that the eigenvalues are arranged in increasing order, allowing for their multiplicity, and that the corresponding eigenelements $\{u_i\}$ are orthonormalized in the energy space H_A of the operator A : $(u_i, u_j)_A = \delta_{ij}$, $i, j = 1, 2, \dots$ (δ_{ij} is the Kronecker delta). We make similar stipulations about the eigenvalues and eigenelements of any pdo with discrete spectrum that may be encountered below.

Zh. vychisl. Mat. mat. Fiz.*, **17, 1, 24–41, 1977.

Assume that there is a self-adjoint pdo A_0 in H with known eigenvalues $\{\lambda_i^0\}$ and eigenelements $\{u_i^0\}$, which is semi-similar to the operator A (i.e., the energy spaces H_A and H_{A_0} consist of the same elements, see [1]), and is connected with it by the relation

$$(u, v)_A = (u, v)_{A_0} + (Su, Sv)_1 \quad \forall u, v \in H_A, \quad (1.1)$$

where S is an operator from H into a Hilbert space H_1 with scalar product $(\cdot, \cdot)_1$ and norm $|\cdot|_1$.

Our problem can then be solved by the Bazley-Fox method (see e.g., [2]), which approximates the required eigenvalues from below. For this, we have to choose in H_1 a sequence of finite-dimensional subspaces $\{W_n\}$, asymptotically dense in H_1 , and after arbitrarily fixing n , we have to evaluate the eigenvalues $\{\lambda_i^n\}$ of the pdo A_n , generated in H by the closed bilinear form $(u, v)_{A_0} + (O_n Su, Sv)_1$, where O_n is the orthogonal projector in H_1 onto W_n . Notice that, to evaluate the $\{\lambda_i^n\}$ we do not need a knowledge of the explicit form of the operator A_n ; the eigenvalues $\{\lambda_i^n\}$ are fully defined by the above bilinear form, with the aid of which the eigenvalue problem for the operator A_n can be written as the identity

$$(u, v)_{A_0} + (O_n Su, Sv)_1 = \lambda(u, v), \quad 0 \neq u \in H_{A_0}, \quad \forall v \in H_{A_0}. \quad (1.2)$$

On varying n , we obtain for each eigenvalue λ_i a sequence λ_i^n , $n=1, 2, \dots$, convergent from below to λ_i . The rate of this convergence is estimated in [3]. It was shown there that, for sufficiently large values of n ,

$$\lambda_i - \lambda_i^n \leq C_1 \sum_{j=0}^{\infty} |(I_1 - O_n) S u_{s+j}|_1^2, \quad (1.3)$$

where I_1 is the identity operator in H_1 , $s+1$ is the multiplicity of the eigenvalue λ_i , s is the least number for which $\lambda_s = \lambda_i$, and C_1 is a positive constant, independent of n .

If we restrict ourselves to the assumptions made above, it becomes necessary to solve complicated transcendental equations in order to determine the eigenvalues λ_i^n , $i=1, 2, \dots$. The computations can be simplified by imposing extra conditions on the subspaces W_n . The condition was originally pointed out by Bazley and Fox [4], and can be stated as

$$\begin{aligned} &\text{whatever the } n=1, 2, \dots, \text{ there exists a number } N(n), \\ &\text{such that } Su_i^0 \perp W_n \text{ for } i > N(n). \end{aligned} \quad (1.4)$$

In this case, the eigenvalues of the problem (1.2) can be found by solving an algebraic eigenvalue problem.

Experience shows that condition (1.4) is extremely rigid, and not many types of problem have as yet been found in which it is satisfied. Moreover, our studies of convergence for some of these problems (see e.g., [5]) have revealed that the rate of convergence of λ_i^n to λ_i may be extremely slow.

In addition to the method considered, involving a special choice of test spaces, Bazley and Fox [4, 6] found another means for overcoming the difficulties that arise when solving the intermediate problem (1.2). Their new idea was to solve problem (1.2) itself approximately, while still retaining the main aim of obtaining lower bounds for the eigenvalues λ_i . While realization of this idea demands certain restrictions, these are much less rigid than condition (1.4). The restrictions are as follows:

1) $\overline{D(S^*)} = H_1$, 2) $W_n \subset D(S^*)$, $n=1, 2, \dots$, where S^* is the operator adjoint to S , and the bar denotes the closure operation. Restriction 1) follows from restriction 2) and the condition made at the start, that the sequence $\{W_n\}$ be asymptotically dense in H_1 . The need for condition 2) is clear from the type of problem which Bazley and Fox proposed to solve instead of (1.2):

$$\lambda_{m+1}^0(u, v) + \sum_{j=1}^m (\lambda_j^0 - \lambda_{m+1}^0)(u, v_j^0)(v_j^0, v) + (\overline{S^* O_n S} u, v) = \lambda(u, v), \quad 0 \neq u \in H, \quad \forall v \in H. \quad (1.5)$$

Here, $v_i^0 = (\lambda_i^0)^{1/2} u_i^0$, $i=1, 2, \dots$, are the eigenelements of the operator A_0 , orthonormalized in H .

Problem (1.5) is obtained from problem (1.2) by replacing the bilinear form $(u, v)_{A_0}$ by the bilinear form, bounded in H ,

$$\sum_{j=1}^m (\lambda_j^0 - \lambda_{m+1}^0)(u, v_j^0)(v_j^0, v) + \lambda_{m+1}^0(u, v), \quad (1.6)$$

and the form $(O_n S u, S v)_1$ by the form $(\overline{S^* O_n S} u, v)$, bounded in H . The second replacement is in a sense the identity transformation. For, by condition 2), $(O_n S u, S v)_1 = (S^* O_n S u, v)$ for all $u, v \in D(S)$, so that the form $(\overline{S^* O_n S} u, v)$ is an extension of the form $(O_n S u, S v)_1$. It is clear that this single replacement does not change problem (1.2). The essence of the Bazley-Fox device lies in the first replacement. As a result of it, a problem is obtained, the eigenvalues of which are lower bounds as before for the eigenvalues λ_j , but they can be computed by solving an algebraic eigenvalue problem. The first of the above-mentioned properties of problem (1.5) follows from the fact that the form (1.6) is less than the form $(u, v)_{A_0}$. To the form (1.6) there corresponds in H a symmetric bounded pdo $A_0^{(m)}$:

$$A_0^{(m)} u = \sum_{j=1}^m (\lambda_j^0 - \lambda_{m+1}^0)(u, v_j^0) v_j^0 + \lambda_{m+1}^0 u,$$

which is called the m -th order truncation of the operator A_0 .

This name originates from the method of obtaining the operator $A_0^{(m)}$: in the spectral decomposition

$$A_0 u = \sum_{j=1}^{\infty} \lambda_j^0 (u, v_j^0) v_j^0$$

of the operator A_0 in the space H , the eigenvalues $\lambda_{m+1}^0, \lambda_{m+2}^0, \dots$ have to be replaced by the same eigenvalue λ_{m+1}^0 . The form (1.6), which can now be written as $(A_0^{(m)} u, v)$, is called the m -th order truncation of the form $(u, v)_{A_0}$, while the method of replacing problem (1.2) by problem (1.5) is known as the truncation method.

The condition 2) enables us to write problem (1.5) in the operator form in the space H :

$$A_0^{(m)} u + \overline{S^* O_n S} u = \lambda u.$$

It can easily be seen, by analyzing the structure of the operator $A_0^{(m)} + \overline{S^* O_n S}$, that determination of its eigenvalues amounts to a problem of linear algebra.

While application of the truncation method rarely presents serious difficulties, it does not

always give good results. Of course the poor results stem from the slow convergence of the method; but it is hard to say what causes this slow convergence, since we know so little as yet about the rate of convergence of the eigenvalues λ_i^{nm} of the problem (1.5) to the exact eigenvalues λ_i .

In [6], Bazley and Fox obtained the estimate

$$\lambda_i^n - \lambda_i^{nm} \leq C_2(n) (\lambda_{m+1}^0)^{-1}. \quad (1.7)$$

They made the assumption that $A = A_0 + B$, where B is a symmetric positive operator, and they took relation (1.1) in the form

$$(u, v)_A = (u, v)_{A_0} + (u, v)_B,$$

so that $H_1 = H_B$, while S is equal to the identity operator considered as an operator from H into H_B . Bazley and Fox did not investigate the dependence of the constant $C_2(n)$ on n . Weinelt attempted to explain it in [7]. Making the additional assumption that W_n is the same as the linear hull of the elements u_1^0, \dots, u_n^0 , and that the operator B is positive definite and comparable in force with the operator A_0^β , $0 \leq \beta \leq 1$, i.e.,

$$|Bu| \leq C_3 |A_0^\beta u| \quad \forall u \in D(A_0^\beta), \quad C_3 = \text{const} > 0,$$

Weinelt obtained for $C_2(n)$ the estimate

$$C_2(n) \leq C_4 (\lambda_n^0)^{2\beta} \quad (1.8)$$

with a constant C_4 which is independent of n . This is all that is known about the convergence of λ_i^{nm} to λ_i^n for fixed n .

An idea of the convergence of λ_i^{nm} to λ_i can be gained by combining (1.7), (1.8) with (1.3). For instance, if B is bounded ($\beta = 0$), it follows from (1.3), (1.7), and (1.8) that the double sequence λ_i^{nm} , $n, m = 1, 2, \dots$, is convergent to λ_i , at a rate not less than

$$C_4 (\lambda_{m+1}^0)^{-1} + C_1 \sum_{j=0}^{\infty} |(E - O_n) u_{n+j}|_B^2,$$

where E is the identity operator in H . Unfortunately, this case is not typical in practice. And in the case of an unbounded operator B , the expressions in question give no satisfactory estimate of the rate of convergence of λ_i^{nm} to λ_i . If the asymptotic behaviour n^σ , $\sigma > 0$, of the eigenvalues $\{\lambda_n^0\}$, is known, then (1.3), (1.7), and (1.8) can be used to extract from the double sequence λ_i^{nm} , $n, m = 1, 2, \dots$, the ordinary sequences $\lambda_i^{n, m(n)}$, $n = 1, 2, \dots$, and to estimate their rate of convergence to λ_i ; the estimates thereby obtained have a low order.

Recall that everything just said about the convergence of λ_i^{nm} to λ_i only holds under the above special assumptions about the operators A, A_0 and the test subspaces W_n . In the general situation considered at the start of the section, nothing is known about the convergence of λ_i^{nm} to λ_i .

This present state of affairs in the Bazley-Fox method compels us to look for new ways of realizing the basic idea of the method, concerning approximate solution of the intermediate problem (1.2).

A new way of simplifying problem (1.2) is described in the present paper. It differs from the method of truncations mainly in the fact that we replace the bilinear form (u, v) in the identity

(1.2) by a larger bilinear form, whereas Bazley and Fox replaced the form $(u, v)_{A_0}$ in it by a smaller bilinear form. The problem that then arises again has the "intermediate" property: its eigenvalues give lower bounds for the eigenvalues λ_j . We shall show that the determination of the eigenvalues of the new problem reduces to a problem of linear algebra. Finally, the feature of principal importance is that, under the general assumptions made below, an estimate can be obtained for the rate of convergence of the new approximate eigenvalues to the exact eigenvalues. The efficiency of the estimate is illustrated by an example of a Neumann problem for a two-dimensional second-order elliptic equation.

2. A new truncation procedure in the Bazley-Fox method

Turning to a detailed treatment of the new approximate method for solving problem (1.2), we first observe that it can be used under the same general assumptions as the Bazley-Fox method itself.

The new method amounts to replacing the bilinear form (u, v) in the identity (1.2) by the symmetric bilinear form

$$\sum_{j=1}^m [(\lambda_j^0)^{-1} - (\lambda_{m+1}^0)^{-1}] (u, u_j^0)_{A_0} (u_j^0, v)_{A_0} + (\lambda_{m+1}^0)^{-1} (u, v)_{A_0}. \quad (2.1)$$

The approximate problem which we propose to solve instead of problem (1.2) then has the form

$$\begin{aligned} (u, v)_{A_0} + (O_n S u, S v) &= \lambda \{ (\lambda_{m+1}^0)^{-1} (u, v)_{A_0} \\ &+ \sum_{j=1}^m [(\lambda_j^0)^{-1} - (\lambda_{m+1}^0)^{-1}] (u, u_j^0)_{A_0} (u_j^0, v)_{A_0} \}, \\ 0 \neq u &\in H_{A_0}, \quad \forall v \in H_{A_0}. \end{aligned} \quad (2.2)$$

It is easily shown that the form (u, v) is less than the form (2.1):

$$\begin{aligned} (u, u) &= \sum_{j=1}^{\infty} (\lambda_j^0)^{-1} |(u, u_j^0)_{A_0}|^2 \leq \sum_{j=1}^m (\lambda_j^0)^{-1} |(u, u_j^0)_{A_0}|^2 \\ &+ (\lambda_{m+1}^0)^{-1} \sum_{j=m+1}^{\infty} |(u, u_j^0)_{A_0}|^2 \\ &= \sum_{j=1}^m [(\lambda_j^0)^{-1} - (\lambda_{m+1}^0)^{-1}] |(u, u_j^0)_{A_0}|^2 + (\lambda_{m+1}^0)^{-1} |u|_{A_0}^2. \end{aligned} \quad (2.3)$$

The form (2.1) is obviously positive definite in H_{A_0} , and corresponds in this space to the symmetric bounded pdo $(A_0^{-1})^{(m)}$:

$$(A_0^{-1})^{(m)} u = \sum_{j=1}^m [(\lambda_j^0)^{-1} - (\lambda_{m+1}^0)^{-1}] (u, u_j^0)_{A_0} u_j^0 + (\lambda_{m+1}^0)^{-1} u. \quad (2.4)$$

We aim to emphasize, by the notation $(A_0^{-1})^{(m)}$ that this operator is formally obtained from the operator A_0^{-1} by means of the same procedure as the operator $A_0^{(m)}$ is obtained from the operator A_0 . In fact, for A_0^{-1} we write the spectral representation

$$A_0^{-1}u = \sum_{j=1}^{\infty} (\lambda_j^0)^{-1} (u, u_j^0)_{A_0} u_j^0$$

in the space H_{A_0} , and then we replace all the eigenvalues of A_0^{-1} with numbers $m+1, m+2, \dots$, in this representation by the $(m+1)$ -th eigenvalue $(\lambda_{m+1}^0)^{-1}$. Hence it is natural to call $(A_0^{-1})^{(m)}$ the m -th order truncation of the operator A_0^{-1} , and call the present method the method of inverse operator truncations.

Let us return to problem (2.2). We shall show that it has eigenvalues representing lower bounds of the eigenvalues of the problem (1.2), and we shall give the method of calculating them.

Considering S as an operator from H_{A_0} into H_1 , we introduce the adjoint operator S' , acting from H_1 into H_{A_0} :

$$(Su, w)_1 = (u, S'w)_{A_0} \quad \forall u \in H_{A_0}, \quad \forall w \in H_1.$$

It was shown in [3] that the operator S' is bounded. Using this operator, we write (2.2) as an operator problem in H_{A_0} :

$$(I + S'O_n S)u = \lambda (A_0^{-1})^{(m)} u, \quad 0 \neq u \in H_{A_0}, \quad (2.5)$$

where I is the identity operator in H_{A_0} . We can write the last equation out more fully as

$$u + \sum_{i,j=1}^{d(n)} (u, S'w_i)_{A_0} \alpha_{ij} S'w_j = \lambda (A_0^{-1})^{(m)} u + \sum_{j=1}^{d(n)} [(\lambda_j^0)^{-1} - (\lambda_{m+1}^0)^{-1}] (u, u_j^0)_{A_0} u_j^0, \quad (2.6)$$

where $w_1, \dots, w_{d(n)}$ is the basis in W_n , (α_{ij}) , $i, j=1, 2, \dots, d(n)$, is the inverse matrix to the matrix

$$D = ((w_i, w_j)_1), \quad i, j=1, 2, \dots, d(n). \quad (2.7)$$

Symmetric pdo's in H_{A_0} appear in both sides of Eq. (2.5). It is clear from (2.6) that the subspace V , stretched over $u_1^0, \dots, u_m^0, S'w_1, \dots, S'w_{d(n)}$, reduces these operators. On solving problem (2.5) in the subspace, we obtain a finite number of eigenvalues of finite multiplicity $\mu_1^{n,m} \leq \dots \leq \mu_{r(n)}^{n,m}$, where $r(n)$ denotes the dimensionality of the subspace V . In the orthogonal complement to V , Eq. (2.6) transforms into the equation $u = \lambda (\lambda_{m+1}^0)^{-1} u$, so that its spectrum in this complement consists of the unique eigenvalue λ_{m+1}^0 , whose corresponding eigenelement is the entire subspace $H_{A_0} \ominus V$. The eigenvalues $\mu_i^{n,m}$ exhaust the spectrum of the problem (2.5).

We arrange the eigenvalues of the problem (2.5), not exceeding λ_{m+1}^0 , in increasing order while allowing for their multiplicities: $\lambda_1^{n,m} \leq \lambda_2^{n,m} \leq \dots$. Clearly, starting from some number i , not exceeding $r(n) + 1$, this sequence becomes "stationary": $\lambda_j^{n,m} = \lambda_{m+1}^0$, $j \geq i$. In the case

when λ_{m+1}^0 is the least eigenvalue of problem (2.5), the entire sequence $\lambda_j^{nm}, j=1, 2, \dots$, is stationary: $\lambda_j^{nm} = \lambda_{m+1}^0, j=1, 2, \dots$.

On now applying to problem (1.2), (2.2) the familiar comparison theorems, and recalling (2.3), we obtain the estimates of interest:

$$\lambda_i^{nm} \leq \lambda_i^n \leq \lambda_i, \quad i=1, 2, \dots \quad (2.8)$$

Since $\lambda_{m+1}^0 \rightarrow \infty$ as $m \rightarrow \infty$, and the right-hand sides of the inequalities (2.8) are independent of m , the number λ_{m+1}^0 cannot be the least eigenvalue of the problem (2.2) for sufficiently large m ; for such values of m , there must necessarily be eigenvalues of finite multiplicity of problem (2.2) to the left of λ_{m+1}^0 , the number of which increases without limit as m increases.

In short, we have established that the inverse operator truncation method reduces the determination of lower bounds for the eigenvalues $\{\lambda_i\}$ to the solution of problem (2.6) in finite-dimensional space V . In turn, this problem is equivalent to the matrix problem

$$Ax = \lambda Bx \quad (2.9)$$

with symmetric positive matrices A and B of order $r(n)$.

The matrices A, B depend on the choice of basis in V . We can assume without loss of generality that $u_1^0, \dots, u_m^0, S'w_1, \dots, S'w_{r(n)-m}$ form a basis in V . With this basis, A and B have the block forms

$$A = \left\| \begin{array}{c|c} E + C^*D^{-1}C & K^* + C^*D^{-1}F \\ \hline K + F^*D^{-1}C & M + F^*D^{-1}F \end{array} \right\|, \quad (2.10)$$

$$B = \left\| \begin{array}{c|c} \Lambda_0 & \Lambda_0 K^* \\ \hline K \Lambda_0 & K \Lambda_0 K^* + \lambda_{m+1}^0 M \end{array} \right\|, \quad (2.11)$$

where

$$\begin{aligned} E &= (\delta_{ij}), \quad i, j=1, 2, \dots, m, \quad \Lambda_0 = ((\lambda_i^0)^{-1} \delta_{ij}), \quad i, j=1, 2, \dots, m, \\ C &= ((S'w_i, u_j^0)_{A_0}), \quad i=1, 2, \dots, d(n), \quad j=1, 2, \dots, m, \\ K &= ((S'w_i, u_j^0)_{A_0}), \quad i=1, 2, \dots, r(n)-m, \quad j=1, 2, \dots, m; \\ F &= ((S'w_i, S'w_j)_{A_0}), \quad i=1, 2, \dots, d(n), \quad j=1, 2, \dots, r(n)-m, \\ M &= ((S'w_i, S'w_j)_{A_0}), \quad i, j=1, 2, \dots, r(n)-m, \\ L_0 &= ([(\lambda_i^0)^{-1} - (\lambda_{m+1}^0)^{-1}] \delta_{ij}), \quad i, j=1, 2, \dots, m. \end{aligned} \quad (2.12)$$

Expressions (2.7), (2.9)–(2.12) represent the computational formulae for the method of inverse operator truncations.

3. Convergence and convergence rate estimate

We shall start our study of the convergence of the approximate eigenvalues λ_i^{nm} , $i=1, 2, \dots$, by seeing how they depend on the index m . From the definition of the truncation $(A_0^{-1})^{(m)}$ it is clear that it decreases monotonically with respect to m : $(A_0^{-1})^{(m)} \geq (A_0^{-1})^{(m+1)} \geq A_0^{-1}$, $m=1, 2, \dots$. In view of this, for fixed n every eigenvalue λ_i^{nm} is monotonically increasing with respect to m :

$$\lambda_i^{nm} \leq \lambda_i^{n, m+1}, \quad m=1, 2, \dots$$

Let us show that the sequence λ_i^{nm} , $m=1, 2, \dots$, converges to λ_i^n . For this, we reduce each of problems (1.2) and (2.2) to an eigenvalue problem for a symmetric bounded operator in H_{A_0} . We have already taken a step towards this reduction in the case of the problem (2.2), by replacing the identity (2.2) by the equation (2.5). The same step, for the problem (1.2), leads to the following equation in H_{A_0} :

$$(I + S'O_n S)u = \lambda A_0^{-1}u. \quad (3.1)$$

We introduce the notation $F_n = I + S'O_n S$. The properties of the operators F_n were examined in detail in [3] and they will be used below without reference. If we make the replacement $v = F_n^{-1/2}u$, in (2.5) and (3.1), it becomes obvious that $(\lambda_i^n)^{-1}$, $i=1, 2, \dots$, are the eigenvalues of the asymmetric completely continuous operator $F_n^{-1/2}A_0^{-1}F_n^{-1/2}$ in H_{A_0} , while $(\lambda_i^{nm})^{-1}$, $i=1, 2, \dots$, are the eigenvalues of the symmetric bounded operator $F_n^{-1/2}(A_0^{-1})^{(m)}F_n^{-1/2}$ in the same space.

By a well-known theorem * (see e.g., [1], p. 258),

$$(\lambda_i^{nm})^{-1} - (\lambda_i^n)^{-1} \leq |F_n^{-1/2}((A_0^{-1})^{(m)} - A_0^{-1})F_n^{-1/2}|_{A_0}. \quad (3.2)$$

It is easily shown that

$$|(A_0^{-1})^{(m)} - A_0^{-1}|_{A_0} = (\lambda_{m+1}^0)^{-1}. \quad (3.3)$$

On further recalling that $|F_n^{-1/2}|_{A_0} \leq 1$, we obtain from (3.2) and (3.3):

$$(\lambda_i^{nm})^{-1} - (\lambda_i^n)^{-1} \leq (\lambda_{m+1}^0)^{-1},$$

or alternatively,

$$\lambda_i^n - \lambda_i^{nm} \leq \lambda_i^n \lambda_i^{nm} (\lambda_{m+1}^0)^{-1}, \quad (3.4)$$

which shows that λ_i^{nm} are convergent to λ_i^n as $m \rightarrow \infty$.

On coarsening the inequality (3.4), we obtain an estimate for the rate of convergence of λ_i^{nm} to λ_i^n :

$$\lambda_i^n - \lambda_i^{nm} \leq \lambda_i^2 (\lambda_{m+1}^0)^{-1}, \quad (3.5)$$

*The theorem was proved in [1] for completely continuous operators, but it remains true for the present operators, only one of which is completely continuous.

in which, as distinct from the case of Bazley and Fox's estimate (1.7), the constant on the right-hand side is independent of n .

We can also obtain from (3.4) effective, practically computable estimates for the error introduced by the operation of truncation of the operator A_0^{-1} . In fact, an upper bound is easily obtained for the eigenvalue λ_i by Ritz's method; this bound usually holds before application of the Bazley-Fox method (the latter is in fact used to estimate the error of the Ritz method). Denoting by $\bar{\lambda}_i$ an upper bound for λ_i , computed by the Ritz method, we find from (3.4) that

$$\lambda_i^n - \lambda_i^{nm} \leq \lambda_i^{nm} \bar{\lambda}_i^0 (\lambda_{m+1}^0)^{-1}, \quad (3.6)$$

$$\lambda_i^n - \lambda_i^{nm} \leq \bar{\lambda}_i^2 (\lambda_{m+1}^0)^{-1}. \quad (3.7)$$

The estimate (3.6) is *a posteriori* while the estimate (3.7) is *a priori*.

We now turn to a study of the behaviour of the approximate eigenvalue λ_i^{nm} as a function of the two indices n and m . To be more precise, we shall henceforth regard λ_i^{nm} as a double sequence and examine its convergence regardless of how n and m tend to ∞ . We shall give the same interpretation to the convergence of other objects (elements, operators) encountered below, dependent on the pair n, m .

The convergence of λ_i^{nm} to λ_i is proved in an elementary way on the basis of the inequalities (1.3) and (3.5):

$$\lambda_i - \lambda_i^{nm} \leq \lambda_i^2 (\lambda_{m+1}^0)^{-1} + C_1 \sum_{j=0}^{\infty} |(I_1 - O_n) S u_{n+j}|_1^2. \quad (3.8)$$

At the same time, (3.8) gives an estimate for the rate of convergence of λ_i^{nm} to λ_i . This is not a limiting estimate, however. A better estimate, of higher order with respect to m , can be obtained by comparing problem (2.2) directly with the initial problem

$$(u, v)_{A_0} + (Su, Sv)_1 = \lambda(u, v), \quad 0 \neq u \in H_A, \quad \forall v \in H_A, \quad (3.9)$$

and estimating the error $\lambda_i - \lambda_i^{nm}$ in accordance with the same scheme as was used in [3] for estimating the error of the Bazley-Fox method without truncation.

Let us briefly run over the scheme. We first have to reduce the initial problem (3.9) to the equivalent operator problem in the space H_{A_0} : $(I + S'S)u = \lambda A_0^{-1}u$, then the latter, and the approximate equation (2.5), have to be transformed respectively to

$$u = \lambda F^{-1} A_0^{-1} u, \quad (3.10)$$

$$u = \lambda F_n^{-1} (A_0^{-1})^{(m)} u, \quad (3.11)$$

where $F = I + S'S$. It is easily seen that $F^{-1} A_0^{-1} = A^{-1}$. For brevity, we introduce the simpler notation T_{nm} for the operator $F_n^{-1} (A_0^{-1})^{(m)}$.

When estimating the error $\lambda_i - \lambda_i^{nm}$ an important role is played by the difference

$$R_{nm} = T_{nm} - A^{-1}, \quad (3.12)$$

and in particular, by its two obvious representations

$$R_{nm} = F^{-1} S' (I_1 - O_n) S T_{nm} + F^{-1} [(A_0^{-1})^{(m)} - A_0^{-1}], \quad (3.13)$$

$$R_{nm} = F_n^{-1} S' (I_1 - O_n) S A^{-1} + F_n^{-1} [(A_0^{-1})^{(m)} - A_0^{-1}], \quad (3.14)$$

and the property

$$|R_{nm}|_{A_0} \rightarrow 0 \quad \text{as } n, m \rightarrow \infty, \quad (3.15)$$

which is proved in the same way as in [3].

The operator R_{nm} characterizes the proximity of the exact problem (3.10) to the approximate problem (3.11); naturally, the error $\lambda_i - \lambda_i^{nm}$ also depends on it. The nature of the latter dependence is proved in the same way as in [3]. In fact, we initially obtain, with the aid of (3.10) and (3.11):

$$(I - \lambda_i A^{-1}) u_i^{nm} = (\lambda_i^{nm})^{-1} (\lambda_i^{nm} - \lambda_i) u_i^{nm} + \lambda_i R_{nm} u_i^{nm}, \quad (3.16)$$

where u_i^{nm} is the eigenelement of the problem (3.1) corresponding to the eigenvalue λ_i^{nm} . Then, multiplying (3.16) scalarly in H_A by the element $P_i u_i^{nm}$, we find

$$0 = (\lambda_i^{nm})^{-1} (\lambda_i^{nm} - \lambda_i) (u_i^{nm}, P_i u_i^{nm})_A + \lambda_i (R_{nm} u_i^{nm}, P_i u_i^{nm})_A, \quad (3.17)$$

where P_i is the orthogonal projector in H_A onto the subspace stretched over $u_s, u_{s+1}, \dots, u_{s+k}$. Finally, noting that $(u_i^{nm}, P_i u_i^{nm})_A = |P_i u_i^{nm}|_A^2$, and for brevity, putting $y = |P_i u_i^{nm}|_A^{-1} u_i^{nm}$, we obtain from (3.17) the required expression

$$\lambda_i - \lambda_i^{nm} = \lambda_i \lambda_i^{nm} (R_{nm} y, P_i y)_A. \quad (3.18)$$

Notice that, when obtaining (3.18), we have tacitly assumed that $|P_i u_i^{nm}|_A > 0$. We justify this assumption below, for sufficiently large n and m .

The next "block" in the scheme of arguments in [3] is to isolate the "unimportant" terms in the scalar product $(R_{nm} y, P_i y)_A$. Replacing R_{nm} in accordance with (3.13) and using relation (1.1), we can write $(R_{nm} y, P_i y)_A$ as the sum

$$\begin{aligned} (R_{nm} y, P_i y)_A = & (O^{(n)} S T_{nm} P_i y, O^{(n)} S P_i y)_1 + |((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} P_i y|_{A_0}^2 + \\ & (O^{(n)} S T_{nm} P^{(i)} y, O^{(n)} S P_i y)_1 + |((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} P^{(i)} y|_{A_0}^2, \end{aligned} \quad (3.19)$$

where $O^{(n)} = I_1 - O_n$, $P^{(i)} = I - P_i$.

We obtain the expression for T_{nm} from relations (3.12) and (3.14):

$$T_{nm} = A^{-1} + F_n^{-1} S' O^{(n)} S A^{-1} + F_n^{-1} ((A_0^{-1})^{(m)} - A_0^{-1}),$$

and we substitute this expression into the first term on the right-hand side of Eq. (3.19). After obvious transformations, we obtain

$$\begin{aligned} (R_{nm} y, P_i y)_A = & \lambda_i^{-1} |O^{(n)} S P_i y|_1^2 + |((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} P_i y|_{A_0}^2 + \\ & + \lambda_i^{-1} |F_n^{-1/2} S' O^{(n)} S P_i y|_{A_0}^2 + (S F_n^{-1} ((A_0^{-1})^{(m)} - A_0^{-1}) P_i y, O^{(n)} S P_i y)_1 + \\ & + (O^{(n)} S T_{nm} P^{(i)} y, O^{(n)} S P_i y)_1 + |((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} P^{(i)} y|_{A_0}^2. \end{aligned} \quad (3.20)$$

This is in fact the required expansion of the scalar product $(R_{nm} y, P_i y)_A$ into essential and inessential terms. Let us show that the last three terms in (3.20) are inessential. We shall show, in fact, that each of them has higher order than $\theta_{nm} = |O^{(n)} S P_i y|_1^2 + |((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} P_i y|_{A_0}^2$. For the first term, the proof is easy:

$$\begin{aligned}
& |(SF_n^{-1}((A_0^{-1})^{(m)} - A_0^{-1})P_i y, O^{(n)}SP_i y)_1| \\
& \leq |S|_{01} |(A_0^{-1})^{(m)} - A_0^{-1}|_{A_0}^{1/2} |((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} P_i y|_{A_0} |O^{(n)}SP_i y|_1 \\
& \leq |S|_{01} |(A_0^{-1})^{(m)} - A_0^{-1}|_{A_0}^{1/2} \theta_{nm},
\end{aligned} \tag{3.21}$$

where $|\cdot|_{01}$ denotes the norm of the bounded operator acting from H_{A_0} into H_1 . For the other two terms, the required bound cannot be obtained directly, since they contain the expression $P^{(i)}y$, the connection of which with the quantities $|Q^{(n)}SP_i y|_1$ and $|((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} P_i y|_{A_0}$ is as yet unknown. The connection is established by:

Lemma

The pair $n(i), m(i)$ exists such that, for $n \geq n(i), m \geq m(i)$, we have

$$\begin{aligned}
& |P_i u_i^{nm}|_A > 0, \\
& |P^{(i)}y|_A \leq C_5 (|O^{(n)}SP_i y|_1 + |((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} P_i y|_{A_0}),
\end{aligned} \tag{3.22}$$

where the constant C_5 is independent of n, m , and i .

The proof is the same as the proof of Lemma 8 in [3].

We can now easily obtain the required estimates:

$$|((A_0^{-1})^{(m)} - A_0^{-1})P^{(i)}y, P_i y|_{A_0} \leq 2C_5 |(A_0^{-1})^{(m)} - A_0^{-1}|_{A_0}^{1/2} \theta_{nm}, \tag{3.23}$$

$$|(O^{(n)}ST_{nm}P^{(i)}y, O^{(n)}SP_i y)_1| \leq 2C_5 |O^{(n)}ST_{nm}|_{01} \theta_{nm}. \tag{3.24}$$

Notice that the convergence to zero of the quantity $|O^{(n)}ST_{nm}|_{01}$ follows from (3.12), (3.15), and Lemma 1 of [3]. In short, when estimating the rate of convergence of λ_i^{nm} to λ_i we can neglect the last three terms in (3.20). The third term on the right of (3.20) also has no influence on the order of smallness of the error $\lambda_i - \lambda_i^{nm}$ since

$$0 \leq |F_n^{-1/2} S' O^{(n)} SP_i y|_{A_0}^2 \leq |S|_{01}^2 |O^{(n)} SP_i y|_1^2. \tag{3.25}$$

A more exact result, which follows from (3.18), (3.20), (3.21), and (3.23)–(3.25), can be stated as follows.

Theorem 1

If the sequence of subspaces $\{W_n\}$ is asymptotically dense in H_1 , then, given any $i = 1, 2, \dots$, a pair $n'(i) \geq n(i), m'(i) \geq m(i)$, can be found such that, for $n \geq n'(i), m \geq m'(i)$ we have the two-sided estimate

$$\begin{aligned}
& 0.5\lambda_i |O^{(n)}SP_i y|_1^2 + 0.5\lambda_i^2 |((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} P_i y|_{A_0}^2 \\
& \leq \lambda_i - \lambda_i^{nm} \leq 2C_6 \lambda_i |O^{(n)}SP_i y|_1^2 + 2\lambda_i^2 |((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} P_i y|_{A_0}^2,
\end{aligned} \tag{3.26}$$

where $C_6 = 1 + |S|_{01}^2$.

If we use the same method as in [3], and expand $P_i y$ in eigenelements $u_s, \dots, u_{s+\kappa}$, corresponding to the eigenvalue λ_i , we easily obtain from (3.26) an estimate connecting the error $\lambda_i - \lambda_i^{nm}$ with the errors of approximation of the elements $Su_s, \dots, Su_{s+\kappa}$.

Theorem 2

Let the conditions of Theorem 1 hold. Then, for $m \geq m'(i)$, $n \geq n'(i)$

$$\lambda_i - \lambda_i^{nm} \leq C_7(\lambda_i) \sum_{j=0}^{\infty} (|O^{(n)} S u_{s+j}|_1^2 + |((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} u_{s+j}|_{A_0}^2), \quad (3.27)$$

where the constant $C_7(\lambda_i)$ depends only on λ_i . In the case of a single eigenvalue ($\kappa=0$) the order of the estimate (3.27) cannot be improved.

It is only possible to estimate the order of smallness of the best approximation $|O^{(n)} S u_j|_1$ if we take concrete operators A , S and concrete subspaces W_n . With regard to the quantity $|((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} u_j|_{A_0}$, an estimation is possible without auxiliary assumptions. We introduce the orthogonal projector Q_m in H_{A_0} onto the subspace stretched over u_1^0, \dots, u_m^0 . We know that Q_m is also the orthogonal projector in H . Using the definition of the truncation $(A_0^{-1})^{(m)}$ and the spectral representation in H_{A_0} of the operator A_0^{-1} , we can easily show that

$$((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} = ((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} (I - Q_m).$$

In the light of this equation and the estimate (3.3), we have

$$|((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} u_j|_{A_0} \leq (\lambda_{m+1}^0)^{-1/2} |(I - Q_m) u_j|_{A_0}. \quad (3.28)$$

If it is assumed that $u_j \in D(A_0^{0.5+\beta})$, $j=1, 2, \dots$, for some $\beta > 0$, then instead of (3.28) we can obtain the better estimate

$$|((A_0^{-1})^{(m)} - A_0^{-1})^{1/2} u_j|_{A_0} \leq (\lambda_{m+1}^0)^{-1/2-\beta} |(E - Q_m) A_0^{0.5+\beta} u_j|. \quad (3.29)$$

For, under the condition indicated,

$$\begin{aligned} |(I - Q_m) u_j|_{A_0} &= |A_0^{-\beta} (E - Q_m) A_0^{0.5+\beta} u_j| \leq |A_0^{-\beta} (E - Q_m)| \\ &\times |(E - Q_m) A_0^{0.5+\beta} u_j| = (\lambda_{m+1}^0)^{-\beta} |(E - Q_m) A_0^{0.5+\beta} u_j|, \end{aligned}$$

which, in conjunction with (3.28), gives (3.29). We have thus proved:

Corollary

If the conditions of Theorem 1 hold, $n \geq n'(i)$, $m \geq m'(i)$ and $u_j \in D(A_0^{0.5+\beta})$, $j=1, 2, \dots$, for some $\beta \geq 0$, then

$$\lambda_i - \lambda_i^{nm} \leq C_7(\lambda_i) \sum_{j=0}^{\infty} |O^{(n)} S u_{s+j}|_1^2 + o((\lambda_m^0)^{-1-2\beta}). \quad (3.30)$$

Note. In the next section we shall consider an example in which the test subspaces form a generalized and not an ordinary sequence $\{W_\alpha\}$. The extension of the earlier results to this case is trivial: we simply have to replace the index n in all the expressions by the symbol α .

4. Application to the Neumann problem

The main practical difficulties that arise when using the new truncation operation are connected with the operator S' . We previously had to deal with this operator in [3, 5], though there it played an auxiliary role. Now, the operator S' occurs in the computational expressions (2.12).

It is only rarely that S' can be found in explicit, closed form. It is fortunately not essential to be able to do this. The practical problem concerned with S' may be stated alternatively as: to select the test subspaces W_n (or $W\alpha$) in such a way that the operator S' can be simply evaluated on the elements of W_n ($W\alpha$). While this last problem is again difficult, the important example given below shows that a solution is possible.

Let Ω be a circle or rectangle, and $\partial\Omega$ the boundary of Ω , $\bar{\Omega} = \Omega \cup \partial\Omega$. Given in $\bar{\Omega}$ the functions $a(x_1, x_2)$, $a_{ij}(x_1, x_2)$, $i, j=1, 2$, satisfying the conditions:

$$\begin{aligned} a_{ij} &= a_{ji}, \quad a_{ij} \in C^{k+1}(\bar{\Omega}), \quad a \in C^k(\bar{\Omega}), \quad k \geq 1, \\ \sum_{i,j=1}^2 a_{ij} \xi_i \xi_j &\geq \tau (\xi_1^2 + \xi_2^2), \quad \tau = \text{const} > 1, \quad a(x_1, x_2) > 1. \end{aligned} \quad (4.1)$$

In the space $H = L_2(\Omega)$ we define the self-adjoint pdo A by the relations

$$D(A) = \{u \mid u \in W_2^2(\Omega), \partial u / \partial N \equiv \mathcal{P} \nabla u \nu = 0 \text{ on } \partial\Omega\},$$

$$Au = -\text{div } \mathcal{P} \nabla u + au,$$

where $\mathcal{P} = (a_{ij})$, $i, j=1, 2$; ν is the unit vector (written in the column form, as are all vectors) of the inward normal to $\partial\Omega$. Hence A is the operator of the Neumann problem for a second-order two-dimensional elliptic equation; and its properties are well known.

We specify the operator A_0 by the expressions

$$D(A_0) = \{u \mid u \in W_2^2(\Omega), \partial u / \partial \nu = 0 \text{ on } \partial\Omega\}, \quad A_0 u = -\Delta u + u.$$

We have

$$H_A = W_2^1(\Omega), \quad (u, v)_A = \iint_{\Omega} (\mathcal{P} \nabla u \nabla v + auv) d\Omega,$$

$$H_{A_0} = W_2^1(\Omega), \quad (u, v)_{A_0} = \iint_{\Omega} (\nabla u \nabla v + uv) d\Omega,$$

so that

$$(u, v)_A = (u, v)_{A_0} + \iint_{\Omega} [(\mathcal{P} - \mathcal{E}) \nabla u \nabla v + (a-1)uv] d\Omega, \quad (4.2)$$

where $\mathcal{E} = (\delta_{ij})$, $i, j=1, 2$. If we put $H_1 = L_2(\Omega) \times L_2(\Omega) \times L_2(\Omega)$, $D(S) = W_2^1(\Omega)$, $Su = \mathcal{R}(u_{x_1}, u_{x_2}, u)^T$, where T denotes transposition, $u_{x_i} = \partial u / \partial x_i$,

$$\mathcal{R} = \begin{pmatrix} (\mathcal{P} - \mathcal{E})^{1/2} & 0 \\ 0 & (a-1)^{1/2} \end{pmatrix},$$

then Eq. (4.2) can be written in the form (1.1) and all the conditions for application of the Bazley-Fox method to the problem $Au = \lambda u$ are satisfied.

In order to find the test subspaces in which we are able to compute the values of the operator S' , we have to analyze in more detail the structure of the space H_1 .

Theorem 3

The space $H_1 = L_2(\Omega) \times L_2(\Omega) \times L_2(\Omega)$ can be written as the orthogonal sum of the three subspaces

$$H_1 = X \oplus Y \oplus Z,$$

where

$$X = \{w \mid w = (u_{x_1}, u_{x_2}, u)^T, u \in W_2^1(\Omega)\},$$

$$Y = \{w \mid w = (u_{x_2}, -u_{x_1}, 0)^T, u \in \dot{W}_2^1(\Omega)\},$$

$$Z = \{w \mid w = (u_{x_1}, u_{x_2}, \Delta u)^T, u \in W_2^2(\Omega), \partial u / \partial \nu = 0 \text{ on } \partial \Omega\}.$$

Proof. It is easily shown that X, Y are subspaces. Their orthogonality follows from the well-known theorem on the decomposition of the space $L_2(\Omega) \times L_2(\Omega)$ into an orthogonal sum of subspaces $G = \{g \mid g = (u_{x_1}, u_{x_2})^T, u \in W_2^1(\Omega)\}$ and $J = \{g \mid g = (u_{x_2}, -u_{x_1})^T, u \in \dot{W}_2^1(\Omega)\}$. For, given any pair $\varphi \in X, \psi \in Y$ we have

$$\varphi = (u_{x_1}, u_{x_2}, u)^T, u \in W_2^1(\Omega), \quad \psi = (v_{x_2}, -v_{x_1}, 0)^T, v \in \dot{W}_2^1(\Omega),$$

$$(\varphi, \psi)_1 = \iint_G (u_{x_1} v_{x_2} - u_{x_2} v_{x_1}) d\Omega,$$

and since $(u_{x_1}, u_{x_2})^T \in G, (v_{x_2}, -v_{x_1})^T \in J$, we have $(\varphi, \psi)_1 = 0$.

It remains to show that $(X \oplus Y)^\perp = Z$. We take an arbitrary element $w = (w_1(x_1, x_2), w_2(x_1, x_2), w_3(x_1, x_2))^T \in (X \oplus Y)^\perp$. We have

$$\iint_\Omega (u_{x_1} w_1 + u_{x_2} w_2 + u w_3) d\Omega = 0 \quad \forall u \in W_2^1(\Omega), \quad (4.3)$$

$$\iint_\Omega (v_{x_2} w_1 - v_{x_1} w_2) d\Omega = 0 \quad \forall v \in \dot{W}_2^1(\Omega). \quad (4.4)$$

By the theorem just mentioned, on decomposition of the space $L_2(\Omega) \times L_2(\Omega)$, it follows from (4.4) that $w_1 = z_{x_1}, w_2 = z_{x_2}$, where $z \in W_2^1(\Omega)$. On using this fact in the identity (4.3), we get

$$\iint_\Omega (u_{x_1} z_{x_1} + u_{x_2} z_{x_2} + u w_3) d\Omega = 0 \quad \forall u \in W_2^1(\Omega).$$

This last identity implies that $z(x_1, x_2)$ is the generalized solution of the Neumann problem

$$\Delta z = w_3 \text{ in } \Omega, \quad \partial z / \partial \nu = 0 \text{ on } \partial \Omega. \quad (4.5)$$

Since Ω is a circle or rectangle, Eqs. (4.5) will hold almost everywhere, in Ω and on $\partial \Omega$ respectively (see [8]). The theorem is proved.

The set $\{u \mid u \in W_2^2(\Omega), \partial u / \partial \nu = 0 \text{ on } \partial \Omega\}$ is a subspace of the space $W_2^2(\Omega)$. For brevity we shall henceforth denote it by \mathfrak{R} .

Let us turn to the construction of the test subspaces. They will now depend on the three integer indices l, p, q . The set \mathfrak{B} of all triples $\alpha = (l, p, q)$ will be assumed to be partially ordered

in the natural way: $\alpha \leq \alpha'$, if $l \leq l'$, $p \leq p'$, $q \leq q'$. Since \mathfrak{A} is a directed set, the test subspaces W_α will form a generalized sequence. Roughly speaking, the subspace W_α will be constructed according to the same principle as that on which the space H_1 was constructed in Theorem 3. We choose three sequences of finite-dimensional subspaces $\{L_i\}$, $\{M_i\}$, $\{U_i\}$, lying respectively in $W_2^1(\Omega)$, $\dot{W}_2^1(\Omega)$ and \mathfrak{R} . After fixing an arbitrary triple $\alpha = (l, p, q)$, we define W_α as the collection of all elements w of the type $w = \mathcal{R}^{-1}(\varphi + \psi + \zeta)$, where $\varphi = (v_{x_1}, v_{x_2}, v)^T$, $v \in L_l$, $\psi = (f_{x_2}, -f_{x_1}, 0)^T$, $f \in M_p$, $\zeta = (z_{x_1}, z_{x_2}, \Delta z)^T$, $z \in U_q$.

Let us find $S'w$ for $w \in W_\alpha$. We have

$$(Su, w)_1 = (\mathcal{R}^{-1}Su, \varphi + \psi + \zeta)_1 = \iint_{\Omega} (\nabla u \nabla v + uv) d\Omega \\ + \iint_{\Omega} (u_{x_1} f_{x_2} - u_{x_2} f_{x_1}) d\Omega + \iint_{\Omega} (\nabla u \nabla z + u \Delta z) d\Omega \quad \forall u \in D(S),$$

and since the last two integrals vanish in accordance with Theorem 3, we have $(Su, w)_1 = (u, v)_{L_0}$. But this implies that $S'w = v$. In short, the difficulty arising in applying the inverse operator truncation method can be overcome in practice in the present example.

In order for the approximate to converge to the exact eigenvalues, the generalized sequence $\{W_\alpha\}$ has to be asymptotically dense in H_1 . Sufficient conditions for this are given by:

Theorem 4

If the sequences $\{L_i\}$, $\{M_i\}$, $\{U_i\}$ are asymptotically dense in $W_2^1(\Omega)$, $\dot{W}_2^1(\Omega)$ and \mathfrak{R} , respectively, then the generalized sequence $\{W_\alpha\}$ is asymptotically dense in H_1 .

Proof. Let Π_l and Φ_p be the orthogonal projectors in $W_2^1(\Omega)$ onto L_l and M_p respectively, and let Γ_q be the orthogonal projector in $W_2^2(\Omega)$ onto U_q . We take an arbitrary element $w \in H_1$, multiply it on the left by the matrix \mathcal{R} , and on the basis of Theorem 3, decompose $\mathcal{R}w$ into a sum of three mutually orthogonal terms:

$$\mathcal{R}w = (\mathcal{R}w)_x + (\mathcal{R}w)_y + (\mathcal{R}w)_z, \quad (4.6)$$

where $(\mathcal{R}w)_x = (v_{x_1}, v_{x_2}, v)^T$, $v \in W_2^1(\Omega)$, $(\mathcal{R}w)_y = (f_{x_2}, -f_{x_1}, 0)^T$, $f \in \dot{W}_2^1(\Omega)$, $(\mathcal{R}w)_z = (z_{x_1}, z_{x_2}, \Delta z)^T$, $z \in \mathfrak{R}$. Using the decomposition (4.6), the orthogonality of the elements $(\mathcal{R}w)_x$, $(\mathcal{R}w)_y$, $(\mathcal{R}w)_z$ and the minimal property of the orthogonal projector O_α , we find that

$$|w - O_\alpha w|_1^2 \leq |\mathcal{R}^{-1}|_1^2 (\|v - \Pi_l v\|_{W_2^1(\Omega)}^2 + \|f - \Phi_p f\|_{\dot{W}_2^1(\Omega)}^2 \\ + \|\Delta(z - \Gamma_q z)\|_{L_2(\Omega)}^2 + \|\nabla(z - \Gamma_q z)\|_{L_2(\Omega)}^2). \quad (4.7)$$

Since $\|\Delta u\|_{L_2(\Omega)}^2 + \|\nabla u\|_{L_2(\Omega)}^2 \leq 2\|u\|_{W_2^2(\Omega)}^2$, the theorem now follows from (4.7) and the hypotheses.

Given a concrete choice of the subspaces $\{L_i\}$, $\{M_i\}$, $\{U_i\}$ we can estimate the rate of convergence of the approximate to the exact eigenvalues. We shall confine ourselves to the case when Ω is a circle. It can be assumed without loss of generality that the center of the circle is the origin. As L_i we take the set of polynomials of degree not higher than i with respect to each of the variables x_1, x_2 , while as U_i we take the linear hull of the first i eigenfunctions of the operator A_0 , and we define M_i by the expression

$$M_i = \{f | f = (x_1^2 + x_2^2 - \rho^2)v, v \in L_i\},$$

where ρ is the radius of the circle Ω .

To estimate the rate of convergence, we use the corollary to Theorem 2, after first replacing the index n in it by the multi-index $\alpha = (l, p, q)$ (see note on the corollary), and replacing the eigenvalues λ_m^0 by their asymptotic form in m . The inequality (3.30) then takes the form

$$\lambda_i - \lambda_i^{\alpha m} \leq C_7(\lambda_i) \sum_{j=0}^{\infty} |(I_1 - O_\alpha) S u_{s+j}|_1^2 + o(m^{-1-2\beta}). \quad (4.8)$$

The inequality (4.8) holds for the values of β for which $u_j \in D(A_0^{0.5+\beta})$, $j=1, 2, \dots$. In the case of a circle, it follows from the assumptions (4.1) that $u_j \in W_2^{2+k}(\Omega)$ (see e.g., [9]). Noting this, and the results of [10], we can assert that $\beta = 0.25 - 0.5\epsilon$, where $\epsilon > 0$ and is arbitrarily small, so that the second term on the right of (4.8) has order $o(m^{-1.5+\epsilon})$.

Let us find estimates for the quantities $|(I_1 - O_\alpha) S u_j|_1^2$, $j=1, 2, \dots$. We write the element $\mathcal{R} S u_j$ as the sum of its projections onto X, Y, Z :

$$\mathcal{R} S u_j = (\mathcal{R} S u_j)_X + (\mathcal{R} S u_j)_Y + (\mathcal{R} S u_j)_Z. \quad (4.9)$$

Let v_j, f_j, z_j be elements of $W_2^1(\Omega)$, $\dot{W}_2^1(\Omega)$ and \mathfrak{R} , respectively, realizing the projections in question, i.e., $(\mathcal{R} S u_j)_X = (v_{jx_1}, v_{jx_2}, v_j)^T$, $(\mathcal{R} S u_j)_Y = (f_{jx_2}, -f_{jx_1}, 0)^T$, $(\mathcal{R} S u_j)_Z = (z_{jx_1}, z_{jx_2}, \Delta z_j)^T$. Proceeding in the same way as when obtaining (4.7), we obtain

$$\begin{aligned} |(I_1 - O_\alpha) S u_j|_1^2 &\leq 2|\mathcal{R}^{-1}|_1^2 (\|v_j - \Pi_l v_j\|_{W_2^1(\Omega)}^2 \\ &+ \|f_j - \Phi_p f_j\|_{W_2^1(\Omega)}^2 + \|z_j - \Gamma_q z_j\|_{W_2^2(\Omega)}^2). \end{aligned} \quad (4.10)$$

Estimation of the quantities $\|v_j - \Pi_l v_j\|_{W_2^1(\Omega)}$, $\|f_j - \Phi_p f_j\|_{W_2^1(\Omega)}$ and $\|z_j - \Gamma_q z_j\|_{W_2^2(\Omega)}$ is a familiar problem in approximation theory and has in fact been solved. The answer given by this theory depends on the differential and integral properties of the functions v_j, f_j , and z_j . Using the decomposition (4.9), the conditions (4.1) and the relation $u_j \in W_2^{2+k}(\Omega)$, it is easily shown that $v_j, f_j, z_j \in W_2^{2+k}(\Omega)$. We then obtain directly from the results of [11, 12] the estimates

$$\|v_j - \Pi_l v_j\|_{W_2^1(\Omega)} = O(l^{-h-1}), \quad (4.11)$$

$$\|f_j - \Phi_p f_j\|_{W_2^1(\Omega)} = O(p^{-h-1}). \quad (4.12)$$

We shall initially estimate the quantity $\|z_j - \Gamma_q z_j\|_{W_2^2(\Omega)}$ by using the minimal property of the orthogonal projector Γ_q and the equivalence of the norms $\|u\|_{W_2^2(\Omega)}$ and $\|A_0 u\|_{L_2(\Omega)}$ (see [8]):

$$\begin{aligned} \|z_j - \Gamma_q z_j\|_{W_2^2(\Omega)} &\leq \left\| z_j - \sum_{i=1}^q (z_j, v_i^0) v_i^0 \right\|_{W_2^2(\Omega)} \\ &\leq C_8 \left\| A_0 z_j - \sum_{i=1}^q (A_0 z_j, v_i^0) v_i^0 \right\|_{L_2(\Omega)}, \end{aligned} \quad (4.13)$$

then we apply Theorem 2 of [13] to the term closing the chain of inequalities (4.13). The result thus obtained will depend on the value of the parameter σ , for which the relation $A_0 z_j \in D(A_0^0)$ holds. We proved above that $z_j \in W_2^{2+k}(\Omega)$. In accordance with [10], this ensures that $A_0 z_j \in D(A_0^{0.5})$ for $k=1$, and $A_0 z_j \in D(A_0^{0.75-0.5\epsilon})$ for $k>1$. For these cases, Theorem 2 of [13] gives respectively

$$\left\| A_0 z_j - \sum_{i=1}^q (A_0 z_j, v_i^0) v_i^0 \right\|_{L_2(\Omega)} = o(q^{-0.5}), \quad (4.14)$$

$$\left\| A_0 z_j - \sum_{i=1}^q (A_0 z_j, v_i^0) v_i^0 \right\|_{L_2(\Omega)} = o(q^{-0.75+0.5\varepsilon}). \quad (4.15)$$

Combining the estimates (4.8), (4.10)–(4.15), we finally get

$$\lambda_i - \lambda_i^{\text{asm}} = O(l^{-2h-2}) + O(p^{-2h-2}) + o(q^{-1-\sigma(h)+\varepsilon}) + o(m^{-1.5+\varepsilon}), \quad (4.16)$$

where $\sigma(1)=0$, $\sigma(k)=0.5$ for $k>1$.

It should be mentioned that the estimate (4.16) is better than the estimate obtained for the same problem in [5], where the Bazley–Fox method is used with a special choice of test spaces.

Translated by D. E. Brown

REFERENCES

1. MIKHLIN, S. G., *Numerical realization of variational methods* (Chislennaya realizatsiya variatsionnykh metodov), Nauka, Moscow, 1966.
2. GOULD, S. H., *Variational methods for eigenvalue problems: An introduction to the Weinstein method of intermediate problems*, U of Toronto Press, 1966.
3. POZNYAK, L. T., On the convergence of the Bazley–Fox method in the problem of the eigenvalues of one bilinear form with respect to another, *Zh. vychisl. Mat. mat. Fiz.*, **13**, No. 4, 839–853, 1973.
4. BAZLEY, N. W., and FOX, D. W., Lower bounds to eigenvalues using operator decompositions of the form B^*B , *Arch. Ration. Mech. Analysis*, **10**, No. 4, 352–360, 1962.
5. POZNYAK, L. T., Application of the Bazley–Fox method to two-dimensional second-order elliptic equations, *Zh. vychisl. Mat. mat. Fiz.*, **16**, No. 1, 83–101, 1976.
6. BAZLEY, N. W., and FOX, D. W., Truncations in the method of intermediate problems for lower bounds to eigenvalues, *J. Res. NBS*, **65B**, No. 2, 105–111, 1961.
7. WEINELT, W., Über apriori Fehlerabschätzungen der Eigenwertnäherungen beim Bazley–Fox–Verfahren, *Beitr. Numer. Math.*, No. 1, 195–236, 1974.
8. LADYZHENSKAYA, O. A., and URAL'TSEVA, N. N., *Linear and quasi-linear equations of elliptic type* (Lineinye i kvazilineinye uravneniya ellipticheskogo tipa), Nauka, Moscow, 1964.
9. AGMON, S., *Lectures on elliptic boundary value problems*, Princeton, Van Nostrand Math. Studies, 1965.
10. GRISVARD, P., Characterisation de quelques espaces d'interpolation, *Arch. Ration. Mech. Analysis*, **25**, No. 1, 40–63, 1967.
11. IL'IN, V. P., Some inequalities in functional spaces and their application to investigation of the convergence of variational processes, *Tr. Matem. in-ta Akad. Nauk SSSR*, **53**, 64–127, 1959.
12. SHAPOSHNIKOVA, T. O., Asymptotic estimates for convergence of Ritz's method in eigenvalue problems, *Izv. vuzov. Matematika*, No. 6 (121), 86–91, 1972.
13. POZNYAK, L. T., Evaluation of lower bounds for eigenvalues of some ordinary differential equations by the Bazley–Fox method, *Zh. vychisl. Mat. mat. Fiz.*, **14**, No. 4, 873–890, 1974.

STATIONARY STRATEGIES IN DIFFERENTIAL GAMES*

O. A. MALAFEEV

Leningrad

(Received 10 January 1975; revised 8 December 1975)

DIFFERENTIAL games with dependent movements are considered. The class of mixed stationary strategies is shown to include ϵ -equilibrium situations.

The dynamic behaviour of the games considered below is specified by the differential system

$$\dot{x} = f(x, u, v), \quad (1)$$

which satisfies the following conditions:

1) $x \in R^m = X$, where R^m is m -dimensional Euclidean space, $t \in [0, \infty)$, $u \in U \subset R^p$, $v \in V \subset R^q$; U, V are compact sets;

2) f is continuous with respect to (x, u, v) in $R^m \times U \times V$;

3) f satisfies a Lipschitz condition with respect to x ;

4) positive numbers M and M' exist, such that, for any $x \in R^m$, $u \in U$, $v \in V$

$$\|f(x, u, v)\| \leq M + M'\|x\|,$$

where $\|x\|$ is the norm of x ;

5) the sets $F'(x) = \{y \mid y = f(x, u, v), u \in U, v \in V\}$ are convex and closed for all $x \in R^m$.

Definition 1. An admissible control in the set U (or V) is a measurable function $u : [t_0, t_1] \rightarrow R^p$ or $v : [t_0, t_1] \rightarrow R^q$ such that $u(t) \in U$ or $v(t) \in V$ for any $t \in [t_0, t_1]$.

Definition 2. A trajectory of the differential system (1) in $[t_0, t_1]$ is an absolutely continuous function $x : [t_0, t_1] \rightarrow R^m$, for which admissible controls u, v exist, such that $\dot{x}(t) = f(x(t), u(t), v(t))$ almost everywhere in $[t_0, t_1]$; $(x(t_0), t_0)$ is called the start, and $(x(t_1), t_1)$ the end, of the trajectory $x(t)$.

Definition 3. The set $F(x_0, t_0, t_1)$ of vectors $x \in R^m$, for which a trajectory $x(t)$ exists in $[t_0, t_1]$ with the start (x_0, t_0) and the end (x, t_1) , is called the set of attainability of the system (1) from the point (x_0, t_0) in the time $t_1 - t_0$.

Condition 5) is not essential, and is added merely to simplify the treatment, while condition 4) can be replaced by the requirement that the solutions of system (1) be continuable into the interval in which the game is considered; this is well known to involve no loss of generality.

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 42-51, 1977.

Let us mention two propositions concerning $F(x_0, t_0, t)$, required later. The proofs can be found in [1].

Proposition 1. For any $x_0 \in R^m$, $t_0 \leq t \in [0, \infty)$ the set $F(x_0, t_0, t)$ is a non-empty compact subset of R^m , and regarded as a function of $F(x_0, t_0, t) = x_0$ is continuous in aggregate in the Hausdorff metric; and (x_0, t_0, t) .

Proposition 2. The mapping $\pi[x_0, t_0, t_1]$, which associates the pair of controls $u(t), v(t)$ in the interval $[t_0, t_1]$ and at the point x_0 , with a trajectory of system (1), $\pi[x_0, t_0, t_1] : U \times V \rightarrow \tilde{F}(x_0, t_0, t_1)$, is continuous.

Here, $\tilde{F}(x_0, t_0, t_1)$ is the space of trajectories of system (1) starting at the point x_0 , furnished with a uniform metric in the interval $[t_0, t_1]$.

Let $(u_1, v_1) * (u_2, v_2)$ denote the admissible control of system (1) in the interval $[t_0, t_2] = [t_0, t_1] \cup [t_1, t_2]$, the contraction of which into $[t_0, t_1]$ is the same as (u_1, v_1) , and whose contraction into $[t_1, t_2]$ is the same as (u_2, v_2) .

1. The game $\Gamma(x_0, T, U, V)$ starts from the point $x_0 \in X$ at the instant $t_0 = 0$ and ends at the instant $T < \infty$. Let E be the maximizing, and P the minimizing player, of the game; at any instant $t \in [0, T]$ they both know k and the state $x(t)$ of the game at t . Let Σ_T be the set of finite divisions $\sigma = \{t_0 = 0 < t_1^\sigma < \dots < t_{N_\sigma}^\sigma = T\}$ of the interval $[0, T]$.

The strategy φ (q, ψ) of the player P (or E) in the game $\Gamma(x_0, T, U, V)$ is the pair $(\xi, \bar{\varphi})$ or $(\eta, \bar{\psi})$, where $\xi, \eta \in \Sigma_T$, $\bar{\varphi} = \{\varphi_\sigma\}_{\sigma \in \Sigma_T}$ or $\bar{\psi} = \{\psi_\sigma\}_{\sigma \in \Sigma_T}$. Here φ_σ (or ψ_σ) is the strategy of P (or E) in the discrete game $\Gamma^\sigma(x_0, T, U, V)$, determined for the division $t_i \in \sigma$ i.e., the mapping associating the information state of P (or E) at the instant $\sigma \in \Sigma_T$, with the probability measure $\mu_i = \mu(t_i, x(t_i))$ or $v_i = v(t_i, x(t_i))$ in U, V at the position $x(t_i)$.

The terminal pay-off in the game $\Gamma(x_0, T, U, V)$ is specified by the function \bar{H} , which satisfies a Lipschitz condition in X . In the game $\Gamma(x_0, T, U, V)$ the pay-off in the situation $(\varphi, \psi) = ((\xi, \bar{\varphi}), (\eta, \bar{\psi}))$ is the mathematical expectation of the pay-off in the discrete game $\Gamma^\sigma(x_0, T, U, V)$ in the situation $(\varphi_\sigma, \psi_\sigma)$, where $\sigma = \xi \cup \eta$. We denote it by $H(\varphi, \psi)$.

Theorem 1

Given any $\epsilon > 0$ there exists an ϵ -equilibrium situation in the game $\Gamma(x_0, T, U, V)$ for any $x_0 \in X$, $T < \infty$.

Proof. It follows from [2, 3] that, for every sequence $\{\sigma_n\}_{n=1}^\infty$ of divisions of the interval $[0, T]$ such that

$$|\sigma_n| = \max_{1 \leq i \leq N_{\sigma_n}} (t_i - t_{i-1}) \rightarrow 0, \quad n \rightarrow \infty,$$

there exists

$$\lim_{n \rightarrow \infty} \text{Val}(\Gamma^{\sigma_n}(x_0, T, U, V)) = V(x_0, T),$$

which is common for all such sequences.

We specify $\epsilon > 0$ and choose $\alpha > 0$ such that, for every $\sigma \in \Sigma_T$ such that $|\sigma| < \alpha$, we have $|V(\cdot) - \text{Val}(\Gamma^\sigma(\cdot))| < \epsilon$. We put

$$\varphi^e = (\sigma_1, \{\hat{\varphi}_\sigma\}_{\sigma \in \Sigma_T}), \quad \psi^e = (\sigma_2, \{\hat{\psi}_\sigma\}_{\sigma \in \Sigma_T}).$$

Here $|\sigma_i| < \alpha$, $i=1, 2$; $\hat{\varphi}_\sigma$ (or $\hat{\psi}_\sigma$) is the optimal strategy of P (or E) in the game $\Gamma^\sigma(x_0, T, U, V)$. We put $\bar{\sigma} = \sigma_1 \cup \sigma_2$. Then, in view of the choice of $\alpha, \bar{\sigma}$, we have $|V(\cdot) - \text{Val}(\Gamma^{\bar{\sigma}}(\cdot))| < \varepsilon$, while in view of the choice of $\hat{\varphi}_{\bar{\sigma}}, \hat{\psi}_{\bar{\sigma}}$ we have $H^{\bar{\sigma}}(\hat{\varphi}_{\bar{\sigma}}, \hat{\psi}_{\bar{\sigma}}) \leq H^{\bar{\sigma}}(\hat{\varphi}_{\bar{\sigma}}, \hat{\psi}_{\bar{\sigma}}) \leq H^{\bar{\sigma}}(\varphi_{\bar{\sigma}}, \psi_{\bar{\sigma}})$ for any strategies $\varphi_{\bar{\sigma}}, \psi_{\bar{\sigma}}$ of players P and E in the game $\Gamma^{\bar{\sigma}}(x_0, T, U, V)$.

Consequently, for the strategies φ, ψ of players P and E in the game $\Gamma(x_0, T, U, V)$ we have $H(\varphi^e, \psi) - \varepsilon \leq H(\varphi^e, \psi^e) \leq H(\varphi, \psi^e) + \varepsilon$.

Along with the game $\Gamma^\sigma(x_0, T, U, V)$ we shall consider the approximating game $\Gamma^\sigma(x_0, T, U_\varepsilon, V_\varepsilon)$, where $U_\varepsilon, V_\varepsilon$ are finite subsets of the sets U and V respectively.

Let us now recall the definition of a recursive game Γ . It is a finite collection of n antagonistic component games Γ_i , i.e., $\Gamma = \{\Gamma_1, \dots, \Gamma_n\}$, each situation $(\varphi^k, \psi^k) \in \Phi^k \times \Psi^k$, $k=1, 2, \dots, n$, being associated with the generalized pay-off

$$\bar{H}^k(\varphi^k, \psi^k) = p^k e^k + \sum_{j=1}^n q^{kj} \Gamma_j, \quad (2)$$

where

$$p^k, q^{kj} \geq 0, \quad p^k + \sum_j q^{kj} = 1.$$

The generalized pay-off means that the player E obtains from player P the quantity e^k with probability p^k , and the recursive game moves over, with probability q^{kj} , to the new state, i.e., the component Γ_j . By the strategy φ of the player P in the game Γ we mean the sequence $\varphi = \{\varphi_t\}_{t=1}^\infty$, where $\varphi_t = (\varphi_t^1, \dots, \varphi_t^n)$, $\varphi_t^i \in \Phi^i$, $t=1, 2, \dots$, $i=1, 2, \dots, n$, so that, if P is in the component game Γ_k at the instant t , then he employs the strategy φ_t^k . The strategy ψ of the player E is defined in a similar way. The value of the pay-off function at a point of Γ_k in the situation (φ, ψ) is taken to be equal to the mathematical expectation of the pay-off during a random walk in the situation (φ, ψ) from the initial position Γ_k . The pay-off in the situation (φ, ψ) is thus defined as

$$H(\varphi, \psi) = (H^1(\varphi, \psi), \dots, H^n(\varphi, \psi)).$$

We say that the component game Γ_k satisfies the minimax condition if the game obtained by replacement of the generalized pay-off \bar{H}^k by the pay-off

$$\bar{H}^k(\varphi^k, \psi^k, \bar{W}) = p^k e^k + \sum_j q^{kj} W_j,$$

where $\bar{W} \in R^n$, has a minimax solution for any $\bar{W} \in R^n$.

Let $\mathbf{1} = (1, \dots, 1)$. The recursive game Γ will be said to have a solution if a vector $V \in R^n$, exists, such that, for every $\varepsilon > 0$, a strategy $\varphi_\varepsilon \in \Phi$, $\psi_\varepsilon \in \Psi$, exists, such that, for any $\varphi \in \Phi$ or $\psi \in \Psi$, we have

$$H(\varphi, \psi_\varepsilon) - \varepsilon \cdot \mathbf{1} \leq V \leq H(\varphi_\varepsilon, \psi) + \varepsilon \cdot \mathbf{1}.$$

The vector V is called the value of the game Γ , while $\varphi_\varepsilon, \psi_\varepsilon$ are called the ε -optimal strategies of players P and E respectively.

The strategy φ (or ψ) is said to be stationary in the component i if $\varphi_i^i = \varphi_i^i$ or $\psi_i^i = \psi_i^i$ for any i . The strategy φ (or ψ) is stationary, if it is stationary in all the components.

Theorem 2

Every recursive game Γ , whose game components have bounded pay-offs and satisfy the minimax condition, possesses a value; ϵ -optimal stationary strategies exist for the players P and E .

For the proof, see [4].

Lemma 1

Given any $x_0 \in X$, $T < \infty$, $\sigma \in \Sigma_T$ and any finite sets $U_\delta \subset U$, $V_\delta \subset V$ the game $\Gamma^\sigma(x_0, T, U_\delta, V_\delta)$ is recursive, and there exist in it ϵ -equilibrium situations in stationary strategies.

Proof. With every point $x = x_0$,

$$x \in \pi[x(t_i), t_i, t_{i+1}](U_\delta, V_\delta)(t_{i+1}), \quad i=0, 1, \dots, N_\sigma-1,$$

we associate a game component Γ_x as follows. The spaces of player's strategies in the game are U_δ, V_δ . With each situation $(u, v) \in U_\delta \times V_\delta$ is associated the generalized pay-off

$$\bar{H}_x(u, v) = \begin{cases} \bar{H}(x) \cdot 1, & \text{if } x \in F(x(t_{N_\sigma-1}), t_{N_\sigma-1}, T), \\ \Gamma_y \cdot 1 & \text{otherwise,} \\ \text{where } y = \pi[x, t_{i+1}, t_{i+2}](u, v)(t_{i+2}). \end{cases}$$

It is easily seen that relations (2) are satisfied here, and hence the game $\Gamma^\sigma(x_0, T, U_\delta, V_\delta)$ is recursive. In view of the finiteness, every game component Γ_x satisfies the minimax condition, and by Theorem 2, the game $\Gamma^\sigma(x_0, T, U_\delta, V_\delta)$ has a value, while the players P and E have ϵ -optimal stationary strategies; and in view of the finiteness of σ , we have $\epsilon = 0$ here.

Lemma 2

The value function $\text{Val}(\Gamma^\sigma(x_0, T, U_\delta, V_\delta)) = V^\sigma(\cdot)$ is continuous with respect to $x_0 \in X$, $T < \infty$.

The proof is similar to that of Lemma 2 in [5].

Let $\rho_{t_0, T}$ denote a uniform metric in the space of functions, continuous in the set $F(x_0, t_0, T)$ and let $\Gamma^\sigma(x_0, T, U, V, \bar{H})$ denote the game in which the terminal pay-off is specified by the function \bar{H} , continuous in X .

Lemma 3

For any $\epsilon > 0$, there exists $\xi > 0$ such that, if $\rho_{t_0, T}(\bar{H}, \bar{H}') < \xi$, then

$$|V^\sigma(\cdot, \bar{H}) - V^\sigma(\cdot, \bar{H}')| < \epsilon.$$

Proof. Consider the functional equations of the game $\Gamma^\sigma(x_0, T, U, V, \bar{H})$:

$$V^\sigma(x_0, T, \cdot) = \iint_{U \times V} V^\sigma(\pi[x_0, t_0, t_1](u, v)(t_1), T - t_1, \cdot) d\mu_0 \cdot dv_0, \quad (3)$$

..... (cont'd)

$$V^{\sigma}(x_{N_{\sigma}-1}, T-t_{N_{\sigma}-1}, \cdot) = \int \int_{U \times V} \bar{H}(\pi[x_{N_{\sigma}-1}, t_{N_{\sigma}-1}, T](u, v)(T)) d\mu_{N_{\sigma}-1}^* dv_{N_{\sigma}-1}^*.$$

Here, μ_i^*, v_i^* are the optimal probability measures, on which the values are reached in the relevant equations. For the proof, we use induction on n , i.e., on the number of interior points in the division σ . For $n = 0$ the lemma follows directly from the properties of the integral. Now assume that the lemma holds for $N_{\sigma} = n+1$, and let us show that it holds for the case when σ contains $n+1$ interior points. In the set $F(x_0, t_0, t_1)$ we consider the function

$$V^{\sigma_1}(x_1, T-t_1, \cdot), \quad \sigma_1 = \{t_1 < \dots < t_{N_{\sigma}} = T\}.$$

By the inductive hypothesis, for every $\epsilon > 0$ and any $x_1 \in F(x_0, t_0, t_1)$ there exists $\delta(\epsilon, x_1)$, such that, if $\rho_{t_1, T}(\bar{H}, \bar{H}') < \delta(\epsilon, x_1)$, then

$$|V^{\sigma_1}(x_1, T-t_1, \cdot, \bar{H}) - V^{\sigma_1}(x_1, T-t_1, \cdot, \bar{H}')| < \epsilon.$$

We consider the number

$$\delta(\epsilon) = \inf_{x_1} \delta(\epsilon, x_1)$$

and we aim to show that $\delta(\epsilon) > 0$. In fact, if $\delta(\epsilon) = 0$, then there exists a sequence $\{x_i\}_{i=1}^{\infty}$ of points of the compact set $F(x_0, t_0, t_1)$, such that $\delta(x_i) \rightarrow 0$, $i \rightarrow \infty$, while

$$\lim_{i \rightarrow \infty} x_i = x_1^0 \in F(x_0, t_0, t_1).$$

For the point x_1^0 the number $\delta(\epsilon, x_1^0) > 0$ will not exist; but this contradicts the inductive hypothesis. We thus find that, given any $\epsilon > 0$, there exists $\delta > 0$ such that, if $\rho_{t_0, T}(\bar{H}, \bar{H}') < \delta$, then

$$\rho_{t_0, t_1}(V^{\sigma_1}(x_1, T-t_1, \cdot, \bar{H}), V^{\sigma_1}(x_1, T-t_1, \cdot, \bar{H}')) < \epsilon.$$

From the functional equations (3) and the assertion of the lemma in the case $N_{\sigma} = n+1$ we obtain the lemma in the case $N_{\sigma} = n+2$. For, let us specify $\epsilon > 0$. Then, $\delta > 0$ exists, such that, if

$$\rho_{t_0, t_1}(V^{\sigma_1}(x_1, T-t_1, \cdot, \bar{H}), V^{\sigma_1}(x_1, T-t_1, \cdot, \bar{H}')) < \delta,$$

then

$$|V^{\sigma}(x_0, T, \cdot, \bar{H}) - V^{\sigma}(x_0, T, \cdot, \bar{H}')| < \epsilon.$$

And for the given $\delta > 0$, there exists $\eta > 0$ such that, if $\rho_{t_0, T}(\bar{H}, \bar{H}') < \eta$, then

$$\rho_{t_0, t_1}(V^{\sigma_1}(x_1, T-t_1, \cdot, \bar{H}), V^{\sigma}(x_1, T-t_1, \cdot, \bar{H}')) < \delta.$$

Instead of the game $\Gamma(x_0, T, U, V)$ it will be convenient below for us to deal with the similarity defined game $\bar{\Gamma}(x_0, T, U, V)$. Let $K(A)$ be the family of all finite subsets of the set A . Player P 's strategy

$$\varphi = (\xi, U', \{\varphi_{\sigma}(U'', V'')\}_{\sigma \in \Sigma_T, U'' \in K(U), V'' \in K(V)})$$

in this game is specified by the division $\xi \in \Sigma_T$, the set $U' \in K(U)$ and the set of P 's strategies in all the possible games $\Gamma^{\sigma}(x_0, T, U'', V'')$. The player E 's strategy is defined in a similar way:

$$\psi = (\eta, V', \{\psi_{\sigma}(U'', V'')\}_{\sigma \in \Sigma_T, U'' \in K(U), V'' \in K(V)}).$$

The pay-off in the situation (φ, ψ) is defined as follows:

$$H(\varphi, \psi) = H^\sigma(\varphi_\sigma(U', V'), \psi_\sigma(U', V')), \quad \sigma = \xi \cup \eta.$$

Definition 4. The strategy φ (or ψ) of player P (or E) in the game $\bar{\Gamma}(x_0, T, U, V)$ is stationary if all the strategies $\{\varphi_\sigma(U'', V'')\}$ (or $\{\psi_\sigma(U'', V'')\}$), participating in the definition of the strategy φ (or ψ), are stationary in the respective discrete games. We shall denote the optimal player's strategies $\varphi_\sigma^*, \psi_\sigma^*$.

Theorem 3

In the game $\bar{\Gamma}(x_0, T, U, V)$ there exist ϵ -equilibrium situations in stationary strategies for any $x_0 \in X, T < \infty, \epsilon > 0$.

Proof. It is easy to show (see e.g., [6]) that, given any one-step game $\Gamma(A, B)$ with continuous pay-off function on a product of compact metric spaces A and B , there exist, for all $\epsilon > 0$, finite sets $A_\epsilon \in K(A), B_\epsilon \in K(B)$, such that each of the values $\text{Val}(\Gamma(A_\epsilon, B_\epsilon)), \text{Val}(\Gamma(A_\epsilon, B)), \text{Val}(\Gamma(A, B_\epsilon))$ differs from $\text{Val}(\Gamma(A, B))$ by not more than ϵ .

Let us show, by induction on the number of points in the division $\sigma \in \Sigma_T$, that, given any $\epsilon > 0$, there exist finite sets $U_\epsilon \in K(U), V_\epsilon \in K(V)$, such that

$$|V^\sigma(x_0, T, U, V) - V^\sigma(x_0, T, U_\epsilon, V_\epsilon)| < \epsilon.$$

For the case $N_\sigma = 1$, i.e., when the division σ contains no interior points, the game $\Gamma^\sigma(x_0, T, U, V)$ is a one-step game, and the theorem follows at once from the above-mentioned results of [6].

Assume that the theorem holds for games $\Gamma^\sigma(x_0, T, U, V)$ such that σ contains not more than n interior points; we shall show that it then holds for games $\Gamma^{\sigma \cup t'}(x_0, T, U, V)$. Assume that the $(n+1)$ -th point of the division is $t' \in (t_m, t_{m+1})$. Consider in the set $F(x_0, t_0, t')$ the function $V^{\sigma'}(x', T-t', U_\epsilon, V_\epsilon)$, which, by Lemma 2, is continuous. Here $\sigma' = \{t' < t_{m+1} < \dots < T\}$.

By the inductive hypothesis and the continuity of the function $V^{\sigma'}(x', T-t', U_\epsilon, V_\epsilon)$ on the compact set $F(x_0, t_0, t')$, given any $\epsilon > 0$, no matter how small, we can choose finite sets $U_\epsilon \in K(U), V_\epsilon \in K(V)$, in such a way that

$$\rho_{t_0, t'}(V^{\sigma'}(x', T-t', U_\epsilon, V_\epsilon), V^{\sigma'}(x', T-t', U, V)) < \epsilon.$$

By Lemma 3, given any $\epsilon' > 0$, there exists $\epsilon > 0$ such that, if

$$\rho_{t_0, t'}(V^{\sigma'}(x', T-t', U_\epsilon, V_\epsilon), V^{\sigma'}(x', T-t', U, V)) < \epsilon,$$

then

$$|V^\sigma(x_0, T, U, V) - \text{Val}(\Gamma_\epsilon^\sigma(x_0, t', U, V))| < \epsilon', \quad (4)$$

where $\Gamma_\epsilon^\sigma(x_0, t', U, V)$ is the game with pay-off function $V^{\sigma'}(x', T-t', U_\epsilon, V_\epsilon)$, specified on $F(x_0, t_0, t')$.

It follows from the inductive hypothesis that, given any $\epsilon > 0$, there exist $U_\epsilon \in K(U), V_\epsilon \in K(V)$, such that

$$|V^\sigma(x_0, T, U_\epsilon, V_\epsilon) - \text{Val}(\Gamma_\epsilon^\sigma(x_0, T, U, V))| < \epsilon. \quad (5)$$

From relations (4) and (5), given $\epsilon > 0$, there exist $U_\epsilon \in K(U)$, $V_\epsilon \in K(V)$, such that

$$|V^\sigma(x_0, T, U_\epsilon, V_\epsilon) - V^\sigma(x_0, T, U, V)| < \epsilon. \quad (6)$$

It can be shown in a similar way that, given any $\epsilon > 0$, there exist $U_\epsilon \in K(U)$, $V_\epsilon \in K(V)$, such that we have respectively

$$|V^\sigma(x_0, T, U_\epsilon, V) - V^\sigma(x_0, T, U, V)| < \epsilon, \quad (7)$$

$$|V^\sigma(x_0, T, U, V_\epsilon) - V^\sigma(x_0, T, U, V)| < \epsilon. \quad (8)$$

Let us turn to a direct proof of the theorem. We specify the number $\epsilon > 0$. By the inequalities (6)–(8) and Theorem 1, there exist $U_\epsilon \in K(U)$, $V_\epsilon \in K(V)$, σ_ϵ , such that

$$|\lim_{|\sigma| \rightarrow 0} V^\sigma(x_0, T, U, V) - V^{\sigma_\epsilon}(x_0, T, U_\epsilon, V)| < \epsilon,$$

$$|\lim_{|\sigma| \rightarrow 0} V^\sigma(x_0, T, U, V) - V^{\sigma_\epsilon}(x_0, T, U, V_\epsilon)| < \epsilon.$$

In the game $\bar{\Gamma}(x_0, T, U, V)$ we define the strategies $\varphi^\epsilon, \psi^\epsilon$ of players P and E respectively as:

$$\varphi^\epsilon = (\sigma_1, U_\epsilon, \{\varphi_{\sigma'}^*(U', V')\}_{\sigma' \in \Sigma_T, U' \in K(U), V' \in K(V)}),$$

$$\psi^\epsilon = (\sigma_2, V_\epsilon, \{\psi_{\sigma'}^*(U', V')\}_{\sigma' \in \Sigma_T, U' \in K(U), V' \in K(V)}).$$

Here, by Lemma 1, the strategies $\varphi_{\sigma'}^*(\cdot), \psi_{\sigma'}^*(\cdot)$ can be assumed to be stationary. By definition of the strategies $\varphi^\epsilon, \psi^\epsilon$, we have, for all $\varphi \in \Phi, \psi \in \Psi$

$$H(\varphi^\epsilon, \psi) - \epsilon \leq H(\varphi^\epsilon, \psi^\epsilon) \leq H(\varphi, \psi^\epsilon) + \epsilon.$$

2. The discussions of Sec. 1 can be extended in a natural way to the case of time-optimal games with dependent movements in the space X . We shall first construct the approximating multi-step games; by means of the results of [4] we shall prove the existence of equilibrium situations; then finally, we shall prove an existence theorem for a continuous time-optimal game, defined in a similar way to the game $\Gamma(x_0, T, U, V)$ of Sec. 1.

We isolate in X a set M , which we call terminal, and, after fixing a point $x_0 \in X$, we consider the set

$$C = \bigcup_{t \in [0, \infty)} F(x_0, t_0, t).$$

Assume that it is compact. We specify a number $\epsilon > 0$ and then find

$$T_\epsilon = \min_{t \in [0, \infty)} \{t | \hat{\rho}(C_t, X \setminus C) \leq \epsilon\}.$$

Here

$$C_t = \bigcup_{t' \in [0, t]} F(x_0, t_0, t'),$$

and $\hat{\rho}$ is the Hausdorff metric.

We also fix the set $U_\epsilon \in K(U)$, $V_\epsilon \in K(V)$ and a division $\sigma_\epsilon \in \Sigma_{T_\epsilon}$ such that the set

$$D(x_0, \epsilon) = \{\pi[x_i, t_i, t_{i+1}](U_\epsilon, V_\epsilon)(t_{i+1})\}, \quad i=0, 1, \dots, N_{\sigma_\epsilon},$$

forms an ϵ -mesh of the set C_{T_ϵ} . For simplicity, we put here $|t_i - t_{i-1}| = \delta_{\sigma_\epsilon} = \delta$ for $i=1, 2, \dots, N_{\sigma_\epsilon}$.

We now construct the dynamic multi-step game $\Gamma^\epsilon(x_0, U_\epsilon, V_\epsilon)$. It proceeds as follows. At the instant $t_0 = 0$ the two players P and E , knowing the initial position x_0 and the instant t_0 , choose respectively the points $u_0 \in U_\epsilon, v_0 \in V_\epsilon$; as a result of this, the game moves from the state x_0 to the state $x_1 = \pi[x_0, t_0, t_1](u_0, v_0)(t_1)$, etc. At the instant t_i the game is in the state $x_i = x(t_i)$.

If $\rho(x_i, M) \leq \varepsilon$, then the game ends and the player E obtains from P the pay-off t_i .

If $y \in \pi[x(t_j), t_j, t_{j+1}](U_e, V_e)(t_{j+1})$, exists, such that $i > j+1$, $\rho(x_i, y) \leq \varepsilon$, then the state x_i is replaced by the state y .

In the other cases, at the instant t_i , the players P and E , knowing the game state $(t_i, x(t_i))$, choose respectively the points $u_i \in U_e$, $v_i \in V_e$ and as a result the game moves from the state x_i to the state $x_{i+1} = \pi[x_i, t_i, t_{i+1}](u_i, v_i)(t_{i+1})$.

Definition 5. We define the strategy φ (or ψ) of the player P (or E) in the game $\Gamma^e(x_0, U_e, V_e)$ as the mapping which associates the informational state of player P (or E) at the instant t_i with the probability distribution $\mu(x_i, t_i)$ (or $\nu(x_i, t_i)$) at the position $x_i = x(t_i)$ in the set U_e (or V_e).

The pay-off in the game $\Gamma^e(x_0, U_e, V_e)$ in the situation (φ, ψ) is the mathematical expectation of the time of the game.

Definition 6. We define a stochastic univalent game $\bar{\Gamma}$ as the collection $\{\Gamma_i\}_{i=1}^n$ of component games with player's strategy spaces Φ_i, Ψ_i and a generalized pay-off of the form

$$H_i(\varphi_i, \psi_i) = e_i + p_i S + \sum_j q_{ij} \Gamma_j,$$

where $p_i, q_{ij} \geq 0$, $p_i + \sum_j q_{ij} = 1$, and e_j is the non-negative pay-off at the i -th step, obtained by player E regardless of whether the game ends or not; p_i is the probability of termination of the game, and q_{ij} is the probability of the game moving from the component i to the component j (see [4]).

Lemma 4

The game $\Gamma^e(x_0, U_e, V_e)$ is a univalent stochastic game.

Proof. We associate the point $x = x_0$,

$$x \in \pi[x(t_i), t_i, t_{i+1}](U_e, V_e)(t_{i+1}), \quad i = 0, 1, \dots,$$

with the component game Γ_x as follows. The strategy spaces in these games are U_e, V_e . With each pair $(u, v) \in U_e \times V_e$ we associate the generalized pay-off $H_x(u, v) = 1 \cdot S$, if $\rho(x, M) \leq \varepsilon$; $H_x(u, v) = 1 \cdot \Gamma_y$, if there exists $y \in \pi[x(t_j), t_j, t_{j+1}](U_e, V_e)(t_{j+1})$, such that $\rho(x, y) \leq \varepsilon$, $i \leq j+1$; $H_x(u, v) = 1 \cdot \Gamma_y + \delta$ in the remaining cases, where $y = [x, t_{i+1}, t_{i+2}](u, v)(t_{i+2})$.

The lemma can now be proved directly.

Definition 7. The set $\Gamma^* = \{\Gamma_x\}$ of component games of the stochastic univalent game $\bar{\Gamma}$ is said to be a "trap" if, every time that the game hits a component of Γ^* , the player E can guarantee that the game stays in the components of Γ^* for an infinite time, so that he thereby obtains an infinitely large pay-off.

We know from [4] that, if there are no traps in a univalent stochastic game and every component game has a minimax solution, then a solution exists in the game $\bar{\Gamma}$. Hence we obtain the following proposition concerning the game $\Gamma^e(x_0, U_e, V_e)$.

Theorem 4

If, in the game $\Gamma^e(x_0, U_e, V_e)$ the player P has a strategy guaranteeing him finite mathematical expectation of the game time, then a ξ -equilibrium situation will exist in the game for any $\xi > 0$.

Now assume that the value of the game can become infinite; this means that, for any $\alpha > 0$, there exists $\psi_\alpha \in \Psi$, such that $H(\varphi, \psi_\alpha) \geq \alpha$ for any $\varphi \in \Phi$.

Under this assumption, it follows from the definition of the game $\Gamma^e(x_0, U_e, V_e)$ and the results of [4] that:

Theorem 5

Given any $\xi > 0$, an ξ -equilibrium situation exists in the game $\Gamma^e(x_0, U_e, V_e)$

We shall now define the multi-step game $\Gamma^\sigma(x_0, U_e, V_e)$. Here, σ is a division of $[0, \infty)$ containing no limit points. Let Σ_∞ denote the set of all such divisions. In this game, at every instant $t_i \in \sigma$ the values of t_i and $x(t_i)$ are known to both players.

The definition of the strategy φ_σ (or ψ_σ) of player P (or E) in the game is similar to the definition in the game $\Gamma^e(x_0, U_e, V_e)$.

Definition 8. The player P 's strategy φ_σ^* is called successful if, for any strategy ψ_σ of player E , the time of the game $\Gamma^\sigma(x_0, U_e, V_e)$ in the situation $(\varphi_\sigma^*, \psi_\sigma)$ is finite, and moreover,

$$\sup_{(\psi)} t(\varphi_\sigma^*, \psi_\sigma) < \infty.$$

Lemma 5

If, in the game $\Gamma^\sigma(x_0, U_e, V_e)$ a successful strategy φ_σ^* is available to the player P , then ξ -equilibrium situations exist in the game for all $\xi > 0$.

Proof. By hypothesis, the discussion of the game $\Gamma = \Gamma^\sigma(x_0, U_e, V_e)$ can be replaced by a discussion of the game $\Gamma' = \langle t(\cdot, \cdot), \Phi_\sigma^*, \Psi_\sigma \rangle$, where Φ_σ^* is the set of successful strategies of the player P . It is then easily shown that Γ' is a stochastic univalent game, and it follows from [4] that ξ -equilibrium situations exist in it for any $\xi > 0$.

We shall now define the continuous time-optimal game $\Gamma(x_0, U, V)$. In this game, at any instant $t \in [0, \infty)$ the values of t and $x(t)$, and set the set M , are known to both players.

Definition 9. We define the strategy φ of the player P in the game $\Gamma(x_0, U, V)$ as the collection

$$\varphi = (\xi, U', \{\varphi_\sigma(U'', V'')\}_{\sigma \in \Sigma_\infty, U'' \in K(U), V'' \in K(V)}),$$

where $\xi \in \Sigma_\infty$, $U' \in K(U)$, $\{\varphi_\sigma(\cdot)\}_{\sigma, U'', V''}$ is the set of P 's strategies in all possible games $\Gamma^\sigma(x_0, U'', V'')$, $\sigma \in \Sigma_\infty$, $U'' \in K(U)$, $V'' \in K(V)$.

We define player E 's strategy in a similar way:

$$\psi = (\eta, V', \{\psi_\sigma(U'', V'')\}_{\sigma \in \Sigma_\infty, U'' \in K(U), V'' \in K(V)}).$$

The pay-off in the situation (φ, ψ) is defined as follows:

$$H(\varphi, \psi) = H^\sigma(\varphi_\sigma, \psi_\sigma), \quad \sigma = \xi \cup \eta.$$

Definition 10. The strategy φ^* of player P is called successful if, given any E 's strategy ψ ,

$$\sup_{\{\psi\}} t(\varphi^*, \psi) < \infty.$$

The concept of a stationary strategy can be introduced in a similar way to that in Sec. 1.

Now assume that the pay-off function is Lipschitz in the set in which it is finite. Then we have:

Theorem 6

If player P has a successful strategy in the game $\Gamma(x_0, U, V)$ then there are ϵ -equilibrium situations in the game for any $\epsilon > 0$, in stationary strategies.

Proof. Instead of the game $\Gamma(x_0, U, V)$ it is sufficient to consider the game $\Gamma = \langle t(\cdot, \cdot), \Phi^*, \Psi \rangle$. Since, by hypothesis, the pay-off function is Lipschitz, we can regard the game as one with a Lipschitz terminal pay-off function, specified in the terminal set M . But then, all the arguments of Section 1, with minor modifications, are applicable to the game, i.e., all the propositions of Section 1 hold, including the theorem on the existence of an ϵ -equilibrium situation in stationary strategies.

Translated by D. E. Brown

REFERENCES

1. EGGERT, D., and VARAJA, P., Representation of a differential system, *J. Different. Equations*, no. 4, 280-299, 1968.
2. FLEMING, W. H., The convergence problem for differential games, II, *Ann. Math. Studies*, No. 52, 195-210, 1964.
3. FLEMING, W. H., The Cauchy problem for degenerate parabolic equations, *J. Math. Mech.*, 13, 987-1008, 1964.
4. EVERETT, H., Recursive games, *Ann. Math. Studies*, No. 39, 47-78, 1957.
5. MALAFEEV, O. A., on the existence of an ϵ -equilibrium situation in dynamic games with dependent movements. *Zh. vychisl. Mat. mat. Fiz.*, 14, 88-98, 1974.
6. WALD, A., Statistical decision functions, in: *Positional games* (Pozitsionnye igry), Nauka, pp. 300-522, Moscow, 1967.

ON A CLASS OF MULTI-STAGE PROBLEMS OF STOCHASTIC OPTIMAL CONTROL*

E. M. BERKOVICH

Moscow

(Received 27 June 1975)

A CLASS of multi-stage stochastic optimal control problems, involving ordinary differential equations, is described. The replacement of the initial problem by finite-difference analogues is justified. A method for solving the special class of multi-stage problems, linear in the phase variable, is described.

Multi-stage stochastic extremal problems [1, 2] describe familiar situations of decision-making in conditions of imperfect information, in engineering, economics, and other fields of human endeavour. In a number of previous papers (see e.g., [3]), finite difference methods for solving two-stage stochastic optimal control problems have been described and studied. In the present paper, similar methods are considered for a class of multi-stage problems with close similarities to those considered in [4].

1. Formulation of the problem

We consider a controlled dynamic system, whose motion (evolution) is described by ordinary differential equations. We assume that the phase trajectory of the system, and the expenses involved in the chosen control, depend on a random "state of nature" with known probability characteristics, but with a realization which is unknown at the start of the motion. During the motion, additional information arrives, regarding the realizations of certain random parameters, connected with the dynamics of the system. The instants of arrival of the additional information split the total time of the motion into several stages, differing in the information pattern for the control selection. At the first stage no additional information is known and the control is sought as a determinate function of time. When selecting the control applied to the system in the subsequent stages, all the information which has arrived at the start of the stage is taken into account. As an estimate of the expense involved in the chosen control we take the expected value of the target functional under the condition that the controls applied to the system in the subsequent stages are optimal. It is required to choose, under these assumptions, the optimal controls of each stage.

Let us state the problem formally. Given the complete probability space (Ω, Σ, P) . The elements $\omega \in \Omega$ are interpreted as the random "states of nature". The time interval $T_0 \leq t \leq T_N$ of the system motion is assumed to be given. The system state at any instant $t \in [T_0, T_N]$ is described by the m -dimensional vector $x(t)$, while the control applied to the system at this instant is described by the r -dimensional vector $u(t)$. We assume that, given any r -dimensional vector $u = u(t)$, measurable in the interval $[T_0, T_N]$, and given any state of nature $\omega \in \Omega$, a unique phase trajectory $x(t; u, \omega)$, $t \in [T_0, T_N]$, of the system is defined; the trajectory is in fact an m -dimensional vector function, satisfying the equations of motion

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 52-63, 1977.

$$\begin{aligned}\dot{x}(t; u, \omega) &= f(x(t; u, \omega), u(t), t, \omega), \\ t \in [T_0, T_N], \quad x(T_0; u, \omega) &= x_0(\omega),\end{aligned}\tag{1.1}$$

where $f(x, u, t, \omega)$ and $x_0(\omega)$ are given m -dimensional vector functions.

The controls have to be measurable functions of time, whose values belong at every instant to a given set $M \subset R^r$. For a fixed state of nature $\omega \in \Omega$ the expense of the chosen control $u = u(t)$, $t \in [T_0, T_N]$, is estimated by the number $F(x(T_N; u, \omega), \omega)$, where $F(x, \omega)$ is a given scalar function.

The movement time interval is divided by the fixed points $T_0 < T_1 < \dots < T_{N-1} < T_N$ into N stages $\Gamma_i = [T_{i-1}, T_i]$, $i = 1, 2, \dots, N$. The points T_i , $i = 1, 2, \dots, N-1$, define the instants of arrival of the additional information. The information arriving at the instant T_i , $1 \leq i \leq N-1$, is the realization b_i of the random q -dimensional vector quantity, which depends in general on the realized state of nature ω and on the applied control. In particular, b_i may be the result of measurements performed with a random error, on the state of the system at the instant T_i (cf. problems of combining control and observation [5, 6, 7]). On the other hand, b_i may characterise the part of the set Ω in which the realized value ω lies, so that a knowledge of b_i improves the degree of information about the realized state of nature.

In short, the realization of the block random quantity $b^i = (b_1, \dots, b_i)$ is known and can be used in selecting the control, at any $(i+1)$ -th stage, $1 \leq i \leq N-1$.

Denote by U_i , $1 \leq i \leq N$, the set of admissible controls of the i -th stage, i.e., the set of all r -dimensional vector functions $u_i = u_i(t)$, measurable in the interval Γ_i and satisfying the inclusions $u_i(t) \in M$, $t \in \Gamma_i$. We introduce the notation $u^i = (u_1, \dots, u_i)$ for the collection of controls applied in the first i stages, $1 \leq i \leq N$. Such a control will be called admissible if its component controls at each stage are admissible.

Consider the problem of choosing the optimal controls at the individual stages. We start with the last stage Γ_N . The control of this stage is chosen for fixed controls u^{N-1} of the previous stages and fixed information b^{N-1} arriving at the start of the stage. As an estimate of the expense involved in the control u_N of the last stage, we take the expected value of the target functional, i.e., the quantity

$$I_N(u_N; u^{N-1}, b^{N-1}) = E_{\omega|b^{N-1}} F(x(T_N; u_N, \omega), \omega),\tag{1.2}$$

where $u^N = (u^{N-1}, u_N)$, $E_{\omega|b^{N-1}}$ is the operator of conditional mathematical expectation. The problem of choosing the best control of the N -stage for fixed u^{N-1} and b^{N-1} consists in minimizing the functional (1.2) with respect to u_N in the set U_N . The quantity

$$I_N^*(u^{N-1}, b^{N-1}) = \inf_{u_N \in U_N} I_N(u_N; u^{N-1}, b^{N-1})\tag{1.3}$$

is the estimate of the expense involved in the optimal control at the N -th stage for fixed u^{N-1} and b^{N-1} .

We consider any i -th stage Γ_i , $2 \leq i \leq N-1$. When choosing the control of this stage, the controls u^{i-1} of the previous stages are fixed, along with the information b^{i-1} arriving at the start

of the stage. Assume that we know, from the solution of the problem of the $(i+1)$ -th stage, the estimate $I_{i+1}^*(u^i, b^i)$ of the expense involved in the optimal control at the $(i+1)$ -th stage for fixed $u^i = (u^{i-1}, u_i)$, $b^i = (b^{i-1}, b_i)$. Since the realization b_i at the i -th stage is not known, as an estimate of the expense involved in the control u_i for fixed u^{i-1} and b^{i-1} , we take the quantity

$$I_i(u_i; u^{i-1}, b^{i-1}) = E_{b_i | b^{i-1}} I_{i+1}^*(u^i, b^i). \quad (1.4)$$

The optimal control of the i -th stage must minimize the functional (1.4) in the set U_i for fixed u^{i-1} , b^{i-1} . We put

$$I_i^*(u^{i-1}, b^{i-1}) = \inf_{u_i \in U_i} I_i(u_i; u^{i-1}, b^{i-1}), \quad i=2, \dots, N-1. \quad (1.5)$$

In particular, from the solution of the problem at the second stage we find the estimate $I_2^*(u_1, b_1)$ of the expense involved in the control of the first stage u_1 and the information b_1 at the instant T_1 under the proviso that the controls at the subsequent stages be optimal. The problem of choosing the optimal control of the first stage, when no additional information has arrived, consists in minimizing the functional

$$I_1(u_1) = E_{b_1} I_2^*(u_1, b_1) \quad (1.6)$$

in the set U_1 . Here, E_{b_1} is the operator of unconditional mathematical expectation. The number

$$\begin{aligned} I^* &= \inf_{u_1 \in U_1} I_1(u_1) \\ &= \inf_{u_1 \in U_1} E_{b_1} \inf_{u_2 \in U_2} E_{b_2 | b^1} \dots \inf_{u_{N-1} \in U_{N-1}} E_{b_{N-1} | b^{N-2}} \\ &\times \inf_{u_N \in U_N} E_{\omega | b^{N-1}} F(x(T_N; u^N, \omega), \omega) \end{aligned} \quad (1.7)$$

characterizes the mean expense involved in the optimal control at each stage. The collection of these parametric extremal problems for each stage is called the N -stage problem of stochastic optimal control.

The feature of the problem for each stage is that the functional is in general specified implicitly. The problem admits of a variety of generalizations and modifications (see e.g., [4]). Notice in particular that, if the information arriving at certain instants can contain errors as a result of purposive activity of an "opponent", then the operators of mathematical expectation in the problems of the respective stages have to be replaced by lowest upper bound operators (cf. the principle of guaranteed result [8]).

2. Convergence of the difference approximations

Let us construct finite-difference analogues of the N -stage problem of stochastic optimal control. For every sufficiently large integer n , $n \geq n^* = \text{const} \geq N$, we consider a mesh, consonant with the N -stage property, in the interval $[T_0, T_N]$, with base-points $T_0 = t_{n0} < \dots < t_{nn} = T_N$, the points T_i , $i=0, 1, \dots, N$, being included in the base-points: $T_i = t_{ni}$, $i=0, 1, \dots, N$. The mesh base-points divide the interval $[T_0, T_N]$ into subintervals of length $\tau_{nj} = t_{n, j+1} - t_{nj}$, $j=0, 1, \dots, n-1$. The sequence of meshes is assumed to be canonical [9], i.e.,

$$\tau_n = \max_{0 \leq j \leq n-1} \tau_{nj} = O(n^{-1}) \quad \text{as} \quad n \rightarrow \infty.$$

The difference analogue of the control of the i -th stage, $1 \leq i \leq N$, is the r -dimensional mesh vector function $v_{(i)n} = (v_{n, n_i-1}, \dots, v_{n, n_i-1})$, $v_{nh} \in R^r$, $n_{i-1} \leq h \leq n_i-1$. The mesh control of the first i stages can be written in the block form $v_n^i = (v_{(1)n}, \dots, v_{(i)n})$, as a collection of difference controls for each stage.

The difference analogue of the phase trajectory, corresponding to the mesh control $v_n = v_n^N = (v_{n0}, \dots, v_{n, n-1})$ and a fixed state of nature $\omega \in \Omega$, is an m -dimensional mesh vector function, whose components $x_{nj}(v_n, \omega)$, $j=0, 1, \dots, n$, satisfy the equations

$$\begin{aligned} x_{n, j+1}(v_n, \omega) &= x_{nj}(v_n, \omega) + \tau_{nj} f(x_{nj}(v_n, \omega), v_{nj}, t_{nj}, \omega), \\ j &= 0, 1, \dots, n-1, \quad x_{n0}(v_n, \omega) = x_0(\omega). \end{aligned} \quad (2.1)$$

We denote by U_{in} , $1 \leq i \leq N$, the set of admissible mesh controls of the i -th stage, i.e., the mesh vector functions $v_{(i)n} = (v_{n, n_{i-1}}, \dots, v_{n, n_i-1})$, whose components satisfy the inclusions $v_{nh} \in M$, $h = n_{i-1}, \dots, n_i-1$.

The difference analogue of the problem at the N -th stage consists in minimizing, for fixed v_n^{N-1} and b^{N-1} the functional

$$\begin{aligned} I_{Nn}(v_{(N)n}; v_n^{N-1}, b^{N-1}) &= E_{\omega|b^{N-1}} F(x_{Nn}(v_n^N, \omega), \omega), \\ v_n^N &= (v_n^{N-1}, v_{(N)n}), \end{aligned} \quad (2.2)$$

with respect to $v_{(N)n} \in U_{Nn}$. We put

$$I_{Nn}^*(v_n^{N-1}, b^{N-1}) = \inf_{v_{(N)n} \in U_{Nn}} I_{Nn}(v_{(N)n}; v_n^{N-1}, b^{N-1}). \quad (2.3)$$

Assume that the difference problems for the N -th, $(N-1)$ -th, \dots , $(i+1)$ -th stages, $2 \leq i \leq N-1$, have already been defined. The difference analogue of the problem at the i -th stage is the extremal problem on minimization, for fixed v_n^{i-1} and b^{i-1} of the functional

$$\begin{aligned} I_{in}(v_{(i)n}; v_n^{i-1}, b^{i-1}) &= E_{b_i|b^{i-1}} I_{i+1,n}^*(v_n^i, b^i), \\ v_n^i &= (v_n^{i-1}, v_{(i)n}), \quad b^i = (b^{i-1}, b_i), \end{aligned} \quad (2.4)$$

in the set U_{in} . We put

$$I_{in}^*(v_n^{i-1}, b^{i-1}) = \inf_{v_{(i)n} \in U_{in}} I_{in}(v_{(i)n}; v_n^{i-1}, b^{i-1}). \quad (2.5)$$

The difference problem at the 1st stage amounts to minimizing the functional

$$I_{1n}(v_{(1)n}) = E_{b_1} I_{2n}^*(v_{(1)n}, b_1) \quad (2.6)$$

in the set U_{1n} . The collection of the extremal problems for each stage forms a difference N -stage problem of stochastic optimal control. The quantity

$$I_n^* = \inf_{v_{(1)n} \in U_{1n}} I_{1n}(v_{(1)n}) \quad (2.7)$$

is the difference analogue of the mean optimal expense estimate (1.7).

Notice that the construction of the difference analogue of the Cauchy problem (1.1) on the basis of the Euler scheme (2.1) is chosen merely to achieve a clear-cut and convenient treatment. While the use of more exact difference schemes does not alter the fact proved below, of approximation with respect to the functional, it may increase the rate of convergence.

We shall say that the sequence of multi-stage difference problems (2.1)–(2.7) approximates the initial multi-stage problem of stochastic optimal control with respect to the functional (cf. [3, 9]), if $I_n^* \rightarrow I^*$ as $n \rightarrow \infty$. Let us state some assumptions about the initial data of the problem posed in Section 1.

Assumption 1. The set M is a bounded, convex, and closed subset of R^r .

Assumption 2. The function $f(x, u, t, \omega)$ satisfies a Lipschitz condition with respect to x, u, t , uniformly with respect to $\omega \in \Omega$; and for fixed x, u, t , it is, like $x_0(\omega)$, a bounded measurable function of $\omega \in \Omega$.

Assumption 3. The function $F(x, \omega)$ is measurable with respect to $\omega \in \Omega$ a uniformly continuous and bounded function of x in every bounded subset of R^n .

The sufficiency of the conditions stated, for approximation of the initial multi-stage problem with respect to a functional, is proved by:

Theorem 1

Let Assumptions 1–3 hold. Then, the sequence of difference multi-stage problems (2.1)–(2.7) functional-wise approximates the multi-stage problem of stochastic optimal control posed in Section 1.

The proof will be carried out in several steps, involving a number of auxiliary propositions.

Lemma 1

Let Assumptions 1–3 hold. Then, given any $i = 2, 3, \dots, N$, the functional $I_i(u_i; u^{i-1}, b^{i-1})$, defined by Eqs. (1.2)–(1.5), is uniformly continuous in the norm of space $L_2^r([T_0, T_i])$ in the set of admissible controls $u^i = (u^{i-1}, u_i)$ uniformly with respect to b^{i-1} . In addition, the functional $I_1(u_1)$ of (1.6) is uniformly continuous in the norm of $L_2^r(\Gamma_1)$ in the set U_1 .

Proof. Notice first that, by Assumptions 1 and 2, the trajectories $x(t; u, \omega)$, $t \in [T_0, T_N]$, of the problem (1.1) define, uniformly with respect to $\omega \in \Omega$ a uniformly continuous mapping of the set of admissible controls into the space of m -dimensional vector functions $C^m([T_0, T_N])$, continuous in $[T_0, T_N]$. In view of this, and Assumption 3, the functional $I_N(u_N; u^{N-1}, b^{N-1})$ of (1.2) is uniformly continuous in the set of admissible controls $u^N = (u^{N-1}, u_N)$ uniformly with respect to b^{N-1} .

We now use induction. Assume that, for some $i = 2, 3, \dots, N-1$, the functional $I_{i+1}(u_{i+1}; u^i, b^i)$ is uniformly continuous in the set of admissible controls $u^{i+1} = (u^i, u_{i+1})$ uniformly with respect to b^i . We shall show that the functional $I_{i+1}^*(u^i, b^i)$ is then also uniformly continuous in the set of admissible controls u^i , uniformly with respect to b^i . Let u^i and \bar{u}^i be admissible controls

of the first i stages, and let ϵ be an arbitrary positive number. For each b^i there exist controls satisfying the inequalities

$$\begin{aligned} I_{i+1}(u_{i+1}; u^i, b^i) - I_{i+1}^*(u^i, b^i) &< \epsilon/2, \\ I_{i+1}(\bar{u}_{i+1}; \bar{u}^i, b^i) - I_{i+1}^*(\bar{u}^i, b^i) &< \epsilon/2. \end{aligned} \quad (2.8)$$

In addition, in view of the uniform continuity with respect to $u^{i+1} = (u^i, u_{i+1})$ of the functional $I_{i+1}(u_{i+1}; u^i, b^i)$ there exists $\delta > 0$ such that, if

$$\|u^i - \bar{u}^i\|_{L^2([T_0, T_1])} < \delta \quad (2.9)$$

we have, for all b^i , the inequalities

$$\begin{aligned} |I_{i+1}(u_{i+1}; u^i, b^i) - I_{i+1}(u_{i+1}; \bar{u}^i, b^i)| &< \epsilon/2, \\ |I_{i+1}(\bar{u}_{i+1}; u^i, b^i) - I_{i+1}(\bar{u}_{i+1}; \bar{u}^i, b^i)| &< \epsilon/2. \end{aligned} \quad (2.10)$$

It follows from (2.8) and (2.10) that, when (2.9) holds, we have, for all b^i ,

$$|I_{i+1}^*(u^i, b^i) - I_{i+1}^*(\bar{u}^i, b^i)| < \epsilon, \quad (2.11)$$

and this last inequality shows that the functional $I_{i+1}^*(u^i, b^i)$ is uniformly continuous.

Applying the operator of conditional mathematical expectation to (2.11), we can prove the uniform continuity with respect to $u^i = (u^{i-1}, u_i)$ of the functional $I_i(u_i; u^{i-1}, b^{i-1})$, uniform with respect to b^{i-1} . This gives us the first part of the lemma. In particular, the functional $I_2(u_2; u_1, b_1)$ is shown to be uniformly continuous in the set of admissible controls $u^2 = (u_1, u_2)$, the continuity being uniform with respect to b_1 . Hence it follows in turn that the functionals $I_2^*(u_1, b_1)$ and $I_1(u_1)$ are uniformly continuous with respect to $u_1 \in U_1$. Lemma 1 is proved.

For any $i = 1, 2, \dots, N$ and any mesh control $v_{(i)n} \in U_{in}$ we denote by $P_n v_{(i)n}$ its piecewise constant continuation into the interval Γ_i . Obviously, $P_n v_{(i)n} \in U_i$. For any block mesh control $v_n^i = (v_{(1)n}, \dots, v_{(i)n})$ we put $P_n v_n^i = (P_n v_{(1)n}, \dots, P_n v_{(i)n})$. From assumptions 1 and 2 and the difference analogue of Gronwall's lemma, we obtain (cf. [9]).

Lemma 2

Given any $\omega \in \Omega$ and any sequence of admissible mesh controls v_n^N , $n \geq n^*$, we have

$$\max_{0 \leq i \leq n} |x(t_{ni}; P_n v_n^N, \omega) - x_{ni}(v_n^N, \omega)| \rightarrow 0$$

as $n \rightarrow \infty$.

Closeness of the continuous and difference phase trajectories implies a definite degree of closeness between the values of the corresponding functionals.

Lemma 3

For all $i = 2, 3, \dots, N$, any value of b^{i-1} , and any sequence of admissible mesh controls of the first i stages v_n^i , $n \geq n^*$, we have the inequalities

$$\lim_{n \rightarrow \infty} \{I_i(P_n v_{(i)n}; P_n v_n^{i-1}, b^{i-1}) - I_{in}(v_{(i)n}; v_n^{i-1}, b^{i-1})\} \leq 0. \quad (2.12)$$

In addition, given any sequence of mesh controls of the first stage $v_{(1)n} \in U_{1n}$, $n \geq n^*$, we have

$$\overline{\lim}_{n \rightarrow \infty} \{I_1(P_n v_{(1)n}) - I_{1n}(v_{(1)n})\} \leq 0. \quad (2.13)$$

Proof. Notice that, by Assumption 3 and Lemma 2, the inequality (2.12) holds for $i = N$. Let us show that, by virtue of (2.12), we have

$$\overline{\lim}_{n \rightarrow \infty} \{I_i^*(P_n v_n^{i-1}, b^{i-1}) - I_{in}^*(v_n^{i-1}, b^{i-1})\} \leq 0. \quad (2.14)$$

In fact, for any b^{i-1} there exists a sequence of mesh controls $v_{(i)n} \in U_{in}$, $n \geq n^*$, for which

$$\lim_{n \rightarrow \infty} \{I_{in}(v_{(i)n}; v_n^{i-1}, b^{i-1}) - I_{in}^*(v_n^{i-1}, b^{i-1})\} = 0. \quad (2.15)$$

Since $I_{in}^*(v_n^{i-1}, b^{i-1}) \leq I_{in}(v_{(i)n}; v_n^{i-1}, b^{i-1})$, for all n , we obtain (2.14) from (2.12) and (2.15). It follows from (2.14) and Fatou's lemma that relation (2.12) remains valid when i is replaced by $i - 1$. Hence (2.12) is proved for all $i = 2, 3, \dots, N$. Hence, in the light of what has been proved, (2.14) holds for $i = 2$; and in turn, this implies that (2.13) holds. Lemma 3 is proved.

A particular consequence of Lemma 3 is the inequality

$$\overline{\lim}_{n \rightarrow \infty} \{I^* - I_n^*\} \leq 0. \quad (2.16)$$

To prove Theorem 1, it now only remains to show that

$$\overline{\lim}_{n \rightarrow \infty} \{I_n^* - I^*\} \leq 0. \quad (2.17)$$

For the proof, we require two further lemmas. Let us first introduce some notation. Given any continuous function $u_i \in U_i$ we denote by $Q_n u_i = v_{(i)n}$ the mesh control for the i -th stage, representing the "projection onto the mesh" of the control u_i , i.e., $v_{nk} = u_i(t_{nk})$, $k = n_{i-1}, \dots, n_i - 1$, $1 \leq i \leq N$.

Obviously, $Q_n u_i \in U_{in}$. For the continuous control of the first i stages $u^i = (u_1, \dots, u_i)$ we put $Q_n u^i = (Q_n u_1, \dots, Q_n u_i)$. We can prove the following in the same way as in [9]:

Lemma 4

Given any $\omega \in \Omega$ and any continuous admissible control $u = u^N(t)$, $T_0 \leq t \leq T_N$, we have

$$\max_{0 \leq i \leq n} |x_{ni}(Q_n u^N, \omega) - x(t_{ni}; u^N, \omega)| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Using a similar method to that when proving Lemma 3, we can prove from Lemmas 1 and 4:

Lemma 5

For all $i = 2, 3, \dots, N$, any value of b^{i-1} , and any admissible continuous control of the first i stages $u^i = (u_1, \dots, u_i)$ we have

$$\overline{\lim}_{n \rightarrow \infty} \{I_{in}(Q_n u_i; Q_n u^{i-1}, b^{i-1}) - I_i(u_i; u^{i-1}, b^{i-1})\} \leq 0.$$

In addition, for any continuous function $u_1 \in U_1$ we have

$$\overline{\lim}_{n \rightarrow \infty} \{I_{1n}(Q_n u_1) - I_1(u_1)\} \leq 0. \quad (2.18)$$

Since, by Lemma 1, the functional $I_1(u_1)$ is uniformly continuous in the set U_1 , and by assumption 1, the set M is convex and closed, then it can easily be shown, in the same way as in [9], that a sequence of functions, continuous in Γ_1 , exists, minimizing the functional $I_1(u_1)$ in the set U_1 . In view of this and (2.18), we obtain the inequality (2.17), which, jointly with (2.16), is equivalent to the equation $\lim_{n \rightarrow \infty} I_n^* = I^*$, hence Theorem 1 is proved.

Notice that, under certain assumptions about the smoothness of the optimal controls of the initial multi-stage problem, we can guarantee the convergence rate estimate $|I_n^* - I^*| = O(\tau_n)$ as $n \rightarrow \infty$. We can arrange for a higher rate of convergence by using difference analogues, more exact than (2.1), of the Cauchy problem (1.1).

The approximating difference problems enable us to construct minimizing sequences for the initial problems of optimal control for each stage. Particular examples of these sequences are the piecewise constant continuations of the mesh controls, which more and more exactly solve the relevant difference problems. Multi-stage stochastic problems, like determinate optimal control problems, may belong to the class of ill-posed variational problems [10]. If it is necessary to construct sequences, convergent to the optimal controls, the difference approximating problems can be subjected to Tikhonov regularization (cf. [3, 9]).

3. Difference multi-stage problems, linear in the phase variable

We shall consider the special class of multi-stage problems, for which the equation of motion and the target functional depend linearly on the system phase state vector. In addition, we shall assume that the equations of motion for each stage are completely defined by the additional information arriving at the start of the stage. In the case described, the problem for each stage amounts to solving a sequence of relatively simple auxiliary extremal problems.

Let the initial state of the system be a given determinate vector $x_0 = x_{n_0}$. In the present section, the difference mesh with respect to the time axis is assumed to be fixed, and to simplify the notation, we shall omit the auxiliary index n , indicating the number of the mesh base-points, in the phase state and control vectors, and in the functionals and sets.

Assume that the difference analogue of the equations of motion of the system at the i -th stage, $1 \leq i \leq N$, for a fixed value b^{i-1} of the additional information arriving at the start of the stage, is

$$x_{k+1} = A_k(b^{i-1})x_k + B_k(v_k, b^{i-1}), \quad k = n_{i-1}, \dots, n_i - 1. \quad (3.1)$$

Here, $A_k(b^{i-1})$ are given $(m \times m)$ matrices, and $B_k(v_k, b^{i-1})$ are m -dimensional vector functions. To unify the notation, we write b^0 for a fixed (e.g., the zero) vector.

At the first stage, for $n_0 = 0 \leq k < n_1$ the matrices A_k in (3.1) are determinate, while the vectors B_k depend only on the control v_k . At the second stage, for $n_1 \leq k < n_2$ they depend on the

random quantity b_1 , the realization of which is known at this stage, etc. For a fixed initial state $x_{n_{i-1}}$ of the system at the i -th stage, and a chosen control $v_{(i)} = (v_{n_{i-1}}, \dots, v_{n_i-1})$ we can find from Eqs. (3.1) for every vector b^{i-1} the unique difference phase trajectory $x_k = x_k(v_{(i)}, x_{n_{i-1}}, b^{i-1})$, $k = n_{i-1}, \dots, n_i$. The function $F(x, \omega)$, defining the target functional, is also linear in x :

$$F(x, \omega) = (\alpha(\omega), x) + \beta(\omega),$$

where $\alpha(\omega)$ is an m -dimensional vector quantity, and $\beta(\omega)$ is a given scalar.

The special feature of the present situation lies in the fact that the extent of the information of the person choosing the control, at any given stage, about the controls of the previous stages, amounts to exact specification of the phase state of the system at the start of the given stage. For instance, at the last (N -th) stage, the system state $x_{n_{N-1}}$ at the start of this stage is known, and so is the supplementary information b^{N-1} arriving during the motion. The difference problem of the N -th stage amounts to minimizing, with respect to $v_{(N)} \in U_N$ for fixed $x_{n_{N-1}}$ and b^{N-1} , the functional

$$I_N(v_{(N)}; x_{n_{N-1}}, b^{N-1}) = (\alpha_N(b^{N-1}), x_{n_N}) + \beta_N(b^{N-1}), \quad (3.2)$$

where $x_{n_N} = x_{n_N}(v_{(N)}, x_{n_{N-1}}, b^{N-1})$, while the m -dimensional vector function $\alpha_N(b^{N-1})$ and the scalar function $\beta_N(b^{N-1})$ are given by the relations

$$\alpha_N(b^{N-1}) = E_{\omega|b^{N-1}} \alpha(\omega), \quad \beta_N(b^{N-1}) = E_{\omega|b^{N-1}} \beta(\omega). \quad (3.3)$$

We put

$$I_N^*(x_{n_{N-1}}, b^{N-1}) = \inf_{v_{(N)} \in U_N} I_N(v_{(N)}; x_{n_{N-1}}, b^{N-1}).$$

At any i -th stage, $1 \leq i \leq N-1$, the initial state of the system $x_{n_{i-1}}$ and the input additional information b^{i-1} are known. The optimal control of the i -th stage minimizes with respect to $v_{(i)} \in U_i$ for fixed $x_{n_{i-1}}$ and b^{i-1} , the functional

$$I_i(v_{(i)}; x_{n_{i-1}}, b^{i-1}) = E_{b_i|b^{i-1}} I_{i+1}^*(x_{n_i}, b^i), \quad (3.4)$$

where $x_{n_i} = x_{n_i}(v_{(i)}, x_{n_{i-1}}, b^{i-1})$, while $I_{i+1}^*(x_{n_i}, b^i)$ is the lower bound of the functional in the problem of the $(i+1)$ -th stage.

We put

$$I_i^*(x_{n_{i-1}}, b^{i-1}) = \inf_{v_{(i)} \in U_i} I_i(v_{(i)}; x_{n_{i-1}}, b^{i-1}). \quad (3.5)$$

In particular, the number $I^* = I_1^*(x_{n_0}, b^0) = I_1^*(x_0, b^0)$, plays a role similar to that of (2.7), namely, the role of mean estimate of the expense in the multi-stage difference problem.

Let us introduce some notation. For every $i = 1, 2, \dots, N$ and for an arbitrary vector b^{i-1} , we denote by $\lambda_k^{(i)}(b^{i-1}) \in R^m$, $k = n_{i-1}, \dots, n_i$, the solution of the following "conjugate system" of the i -th stage:

$$\lambda_{n_i}^{(i)}(b^{i-1}) = -\alpha_i(b^{i-1}), \quad (3.6)$$

$$\lambda_k^{(i)}(b^{i-1}) = A_k^T(b^{i-1}) \lambda_{k+1}^{(i)}(b^{i-1}), \quad k = n_{i-1}, \dots, n_i-1.$$

Here, T denotes matrix transposition, the m -dimensional vector function $\alpha_i(b^{i-1})$ is given, for $i = N$, by the condition (3.3), and for the other $i = 1, 2, \dots, N-1$, is defined by the recurrence relation

$$\alpha_i(b^{i-1}) = -E_{b_i|b^{i-1}} \lambda_{n_i}^{(i+1)}(b^i), \quad i = 1, 2, \dots, N-1. \quad (3.7)$$

We shall assume that, for every $i = 1, 2, \dots, N$, and any b^{i-1} , there exists a mesh control $v_{(i)}^*(b^{i-1}) = (v_{n_{i-1}}^*, \dots, v_{n_i}^*) \in U_i$, whose components are the solutions of the following supplementary extremal problems:

$$\gamma_k^{(i)}(b^{i-1}) = \max_{v_k \in M} (\lambda_{k+1}^{(i)}(b^{i-1}), B_k(v_k, b^{i-1})), \quad k = n_{i-1}, \dots, n_i - 1. \quad (3.8)$$

We also define the scalar functions $\beta_i(b^{i-1})$, $i = 1, 2, \dots, N$, by the relations

$$\beta_i(b^{i-1}) = E_{b_i|b^{i-1}} \left[\beta_{i+1}(b^i) - \sum_{k=n_i}^{n_{i+1}-1} \gamma_k^{(i+1)}(b^i) \right], \quad i = 1, 2, \dots, N-1, \quad (3.9)$$

where the functions $\gamma_k^{(i)}(b^{i-1})$ are defined by condition (3.8), and $\beta_N(b^{N-1})$ by condition (3.3). It is assumed throughout that the application of the operators of mathematical expectation is valid, though this can be proved under fairly natural assumptions.

In problems for which the arrival of the additional information narrows the domain of possible realizations of the state of nature, the random quantities b_i as a rule have discrete distributions with a finite number of possible realizations. In such cases, the operations of finding the mathematical expectations can be performed in a particularly simple way.

Theorem 2

Under the above assumptions, there exists, in the problem of any i -th stage, $1 \leq i \leq N$, an optimal control which is independent of the controls of the previous stages, and is determined solely by the supplementary information b^{i-1} arriving at the start of the stage. Such an optimal control is, in particular, the mesh function $v_{(i)}^*(b^{i-1})$, whose components solve the auxiliary extremal problems (3.8).

Proof. We shall first show that, if the functional of the i -th stage problem is

$$I_i(v_{(i)}; x_{n_{i-1}} b^{i-1}) = (\alpha_i(b^{i-1}), x_{n_i}) + \beta_i(b^{i-1}), \quad (3.10)$$

where $x_{n_i} = x_{n_i}(v_{(i)}, x_{n_{i-1}}, b_{i-1})$, then its minimum with respect to $v_{(i)} \in U_i$ for fixed $x_{n_{i-1}}$ and b^{i-1} is achieved on the mesh control $v_{(i)}^*(b^{i-1})$. Using (3.6) and (3.1), with arbitrary $v_{(i)} = (v_{n_{i-1}}, \dots, v_{n_i})$ we can transform (3.10) to

$$\begin{aligned} I_i(v_{(i)}; x_{n_{i-1}}, b^{i-1}) &= -(\lambda_{n_{i-1}}^{(i)}(b^{i-1}), x_{n_{i-1}}) \\ &+ \beta_i(b^{i-1}) - \sum_{k=n_{i-1}}^{n_i-1} (\lambda_{k+1}^{(i)}(b^{i-1}), B_k(v_k, b^{i-1})). \end{aligned} \quad (3.11)$$

It in fact follows from (3.11) that the functional (3.10) is minimized on the control $v_{(i)}^*(b^{i-1})$, where, by (3.5) and (3.8),

$$I_i^*(x_{n_{i-1}}, b^{i-1}) = -(\lambda_{n_{i-1}}^{(i)}(b^{i-1}), x_{n_{i-1}}) + \beta_i(b^{i-1}) - \sum_{h=n_{i-1}}^{n_i-1} \gamma_h^{(i)}(b^{i-1}). \quad (3.12)$$

It remains to show that the functional of the i -th stage problem, $1 \leq i \leq N$, in fact has the form (3.10). With $i = N$, (3.10) holds by virtue of (3.2). We then argue by induction. Assume that (3.10) holds for some i , $2 \leq i \leq N$. Then, by (3.4), (3.7), (3.9), and (3.12), the representation (3.10) also holds for the functional of the $(i-1)$ -th stage. Theorem 2 is proved.

In the case when the number of possible realizations of b_i is finite and not unduly large, Theorem 2 provides the basis for the following method of solving the multi-stage problem posed in the present section. In advance of starting the motion of the system, at the stage of processing the *a priori* information, the conjugate systems (3.6) and the auxiliary extremal problems (3.8) are solved. As a result, the controls $v_{(i)}^*(b^{i-1})$ will be constructed for all possible values of b^{i-1} , $i=1, 2, \dots, N$. This preliminary stage can prove to be lengthy and laborious, though not particularly high standards are demanded at this stage concerning the operational properties of the choice of controls. Then, during the motion of the system, the operating side can quickly react to the arrival of the additional information by choosing the previously calculated appropriate control; these tactics prove to be optimal in the mean when an operation is repeated with sufficient frequency.

Translated by D. E. Brown

REFERENCES

1. YUDIN, D. B., *Mathematical methods of control in conditions of imperfect information* (Matemicheskie metody upravleniya v usloviyakh nepolnoi informatsii), Sov. Radio, Moscow, 1974.
2. ERMOL'EV, YU. M., *Methods of stochastic programming* (Metody stokhasticheskogo programmirovaniye), Nauka, Moscow, 1976.
3. BERKOVICH, E. M., On the approximation of two-stage stochastic extremal problems, *Zh. vychisl. Mat. mat. Fiz.*, **11**, No. 5, 1150-1165, 1971.
4. BERKOVICH, E. M., On multi-stage problems of stochastic optimal control, *Izv. Akad. Nauk SSSR, Tekhn. kibernetika*, No. 1, 12-19, 1975.
5. FEL'DBAUM, A. A., *Foundations of the theory of optimal automatic systems* (Osnovy teorii optimal'nykh avtomaticheskikh sistem), Nauka, Moscow, 1966.
6. CHERNOUS'KO, CH. L., Optimization of control and observation process in a dynamic system subject to random disturbances, *Avtomat. telemekhan.*, No. 4, 42-49, 1972.
7. SOLYANIK, A. I., On the optimal combination of pulse control and discrete observation processes in a dynamic system subject to random disturbances, *Kibernetika*, No. 5, 88-98, 1973.
8. GERMEIER, YU. B., *Introduction to operations research theory* (Vvedenie v teoriyu issledovaniya operatsii), Nauka, Moscow, 1971.
9. BUDAK, B. M., BERKOVICH, E. M., and SOLOV'EVA, E. N., On the convergence of difference approximations for optimal control problems, *Zh. vychisl. Mat. mat. Fiz.*, **9**, No. 3, 522-547, 1969.
10. TIKHONOV, A. N., On methods of regularization of optimal control problems, *Dokl. Akad. Nauk SSSR*, **162**, No. 4, 763-765, 1965.

AN EXISTENCE THEOREM IN A MINIMAX CONTROL PROBLEM*

N. S. VASIL'EV

Moscow

(Received 11 June 1975)

A PROBLEM posed by N. N. Moiseev is considered; it can be treated as an application of the principle of the maximum guaranteed result when undetermined factors are present [1]. Necessary conditions for optimality, in the form of a Pontryagin maximum principle, have been discussed on several occasions (see [2-4]). In this connection, it becomes necessary to obtain existence theorems in problems of this kind.

1. The problem

Given a controlled process with a parameter

$$dx/dt = f(x, u(t), v), \quad (1)$$

where $f(x, u, v)$ is a vector function, bounded in the set $X \times P \times Q \subset E^{n+p+q}$; t is the time, $t \geq 0$; $x = (x_1, \dots, x_n)$ is the phase vector, lying in the set X of space E^n ; $u(t) = (u_1(t), \dots, u_p(t))$ is the controlling vector function with values in the set $P \subset E^p$; $v = (v_1, \dots, v_q)$ is a parameter, belonging to the set $Q \subset E^q$.

We fix an arbitrary vector $x^0 \in X$ and an arbitrary positive number T .

Definition. The control $u(t)$ is called admissible if it is Lebesgue measurable in the time interval $[0, T]$, takes values from the set P , and for all parameter values from the set Q , there exists a solution of the given system of differential equations, defined in the interval $[0, T]$ with the initial condition x^0 , the trajectory of which lies in the set X .

Let Ω denote the set of admissible controls. We assume that Ω is not empty. Then the following functional is defined in the set of solutions of the system of differential equations, corresponding to controls of Ω :

$$J(u(t), v) = \int_0^T f^0(x(t), u(t), v) dt, \quad (2)$$

where the integrand is defined in the set $X \times P \times Q$. The question arises as to the conditions in which an admissible control exists, on which the minimum is reached in the expression

$$\inf_{\Omega} \sup_{v \in Q} J(u(t), v).$$

Whatever admissible control is fixed, it is easy to arrange for the functional to reach its maximum with respect to the parameter.

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 64-71, 1977.

Proposition. If Ω is not empty, $X \times P \times Q$ is a compactum in E^{n+p+q} , and the right-hand side of the system and the integrand are continuous with respect to their sets of variables in the set $X \times P \times Q$, then, given any $u(t)$ of the set Ω ,

$$\max_{v \in Q} J(u(t), v)$$

is reached for some value of the parameter.

Proof. We fix an admissible control $u(t)$ and show that the functional is continuous with respect to the parameter in the compactum Q .

Since X is compact, the set of solutions $x_k(t)$, $k=1, 2, \dots$, of the system of differential equations with the parameter v_k , $k=1, 2, \dots$, is uniformly bounded, and in the light of the inequality

$$\begin{aligned} |x_k(t_2) - x_k(t_1)| &\leq \int_{t_1}^{t_2} |f(x_k(t), u(t), v_k)| dt \\ &\leq \max_{X \times P \times Q} |f(x, u, v)| |t_2 - t_1| \end{aligned}$$

for $k=1, 2, \dots$, is equicontinuous. By the Ascoli-Arzelà theorem, the set $\{x_k(t), k=1, 2, \dots\}$ is relatively compact in the space of continuous functions, and it can therefore be assumed that $x_k(t)$ are convergent in a continuous metric to a function $\bar{x}(t)$. Since the set Q is compact, we can assume that the v_k converge to the parameter v , belonging to the set Q .

Since the right-hand side of system (1) is continuous, for all t we have the convergence

$$f_i(x_k(t), u(t), v_k) \rightarrow f_i(\bar{x}(t), u(t), \bar{v}) \quad \text{as } k \rightarrow \infty, i=1, 2, \dots, n,$$

and all the terms of this sequence are uniformly bounded by the constant

$$\max_{X \times P \times Q} |f_i(x, u, v)|, \quad i=1, 2, \dots, n.$$

Then, by Lebesgue's theorem on passage to the limit under the integral sign, we get

$$x_k(t) = x^0 + \int_0^t f(x_k(\tau), u(\tau), v_k) d\tau \rightarrow \bar{x}(t) = x^0 + \int_0^t f(\bar{x}(\tau), u(\tau), \bar{v}) d\tau.$$

Hence $\bar{x}(t)$ is a solution of the system with the parameter \bar{v} .

For a similar reason, the right-hand side of the inequality

$$\begin{aligned} &|J(u(t), v_k) - J(u(t), \bar{v})| \\ &\leq \int_0^t |f^0(x_k(t), u(t), v_k) - f^0(\bar{x}(t), u(t), \bar{v})| dt. \end{aligned}$$

tends to zero. Hence the proposition follows.

Since the class Ω is not in general a compact set, the minimum may not be reached in the expression

$$\inf_{\Omega} \max_Q J(u(t), v)$$

for smooth functions in (1) and (2), and compact bounded sets P and Q .

$$\text{Example. } dx/dt = x(vu(t) + (1-v)(1-u(t))^2), \quad x(0)=1, \\ 0 \leq t \leq 1, \quad P=Q=[0,1], \quad J(u(t), v) = \int_0^1 (x(t) - e^{t/2})^2 dt.$$

The set Ω is the same as the set of Lebesgue-measurable functions, taking values in the interval $[0, 1]$, since the solution may be continued even indefinitely on such controls.

Given any admissible control, the functional reaches its maximum with respect to the parameter. This follows from our proposition, if we note that the phase variable does not leave a compact interval of the straight line R , no matter what the admissible controls or the parameter (see the corollary to the theorem).

Let us show that, in our example, we have

$$\inf_{u \in \Omega} \max_{v \in Q} J(u(t), v) = 0 \quad (3)$$

and that there is no admissible control realizing this.

We take the following sequence of piecewise constant controls, taking only the two alternative values 0 and 1:

$$u_n(t) = \begin{cases} 0, & 0 \leq t < 1/2n, \\ 1, & 1/2n \leq t < 1/n, \\ 0, & 1/n \leq t < 3/2n, \\ \vdots & \vdots \\ 1, & 1 - 1/2n \leq t \leq 1, \end{cases} \\ n=1, 2, \dots$$

We can write the parametrically dependent sequence of solutions of the differential equation $x_n(t, v)$, $n=1, 2, \dots$, corresponding to these controls, in the explicit form

$$x_n(t, v) = \begin{cases} \exp \left[\frac{(1-2v)}{2n} (-l) + (1-v)t \right], & \frac{l}{n} \leq t < \frac{2l+1}{2n}, \\ \exp \left[\frac{(1-2v)}{2n} (l+1) + vt \right], & \frac{2l+1}{2n} \leq t \leq \frac{l+1}{n}, \end{cases} \\ l=0, 1, \dots, n-1,$$

or

$$x_n(t, v) = e^{t/2} h_n(t, v), \quad n=1, 2, \dots, \\ h_n(t, v) = \begin{cases} \exp \left[\frac{1}{2} \left(t - \frac{l}{n} \right) (1-2v) \right], & \frac{l}{n} \leq t < \frac{2l+1}{2n}, \\ \exp \left[\frac{1}{2} \left(\frac{l+1}{n} - t \right) (1-2v) \right], & \frac{2l+1}{2n} \leq t \leq \frac{l+1}{n}, \end{cases} \\ l=0, 1, \dots, n-1.$$

From the inequalities $e^{-1/4n} \leq h_n(t, v) \leq e^{1/4n}$, $n=1, 2, \dots$, there follows the convergence, uniform with respect to time and the parameter, of $x_n(t, v)$ to the function $e^{t/2}$ as $n \rightarrow \infty$, whence we obtain Eq. (3).

Assume that an admissible control $u(t)$ exists, for which

$$\max_{v \in Q} J(u(t), v) = 0.$$

This can only occur under the condition $x(t, v) = e^{t/2}$ for all values of the parameter and of time, where $x(t, v)$ satisfies the equation with the control $u(t)$. Hence, substituting the function $e^{t/2}$ in the differential equation, we arrive finally at the identity

$$e^{t/2}/2 = e^{t/2}(vu(t) + (1-v)(1-u(t))^2),$$

or

$$v[u(t) - 1/2] + (1-v)[(1-u(t))^2 - 1/2] = 0, \quad 0 \leq t \leq 1, \quad 0 \leq v \leq 1.$$

On taking the parameter equal to 0 and 1, we find respectively that the following mutually contradictory equations need to be satisfied:

$$(1-u(t))^2 = 1/2 \quad \text{and} \quad u(t) = 1/2.$$

Note. Given any fixed value of the parameter $v \in [0, 1]$ there exists an admissible control on which is realized

$$\min_{u(t) \in \Omega} J(u(t), v) = 0.$$

The following is an example of such an optimal control:

$$u(v, t) = \begin{cases} \frac{2-3v}{2(1-v)} + \frac{(5v^2-6v+2)^{1/2}}{2(1-v)}, & 1/2 \leq v \leq 1, \\ \frac{2-3v}{2(1-v)} - \frac{(5v^2-6v+2)^{1/2}}{2(1-v)}, & 0 \leq v < 1/2. \end{cases}$$

Let us show that $u(v, t)$ is an admissible control. It is easily shown that the control is non-negative. For $0 \leq v < 1/2$, from the inequalities $2-3v \leq 2(1-v) \leq 2(1-v) + (5v^2-6v+2)^{1/2}$ we obtain $u(v, t) \leq 1$. For $1/2 \leq v \leq 1$, the fact that $u(v, t) \leq 1$ is easily seen from the estimates

$$\begin{aligned} u(1/2, t) &= 1, \quad u(1, t) = 1/2 < 1, \\ \frac{\partial u(v, t)}{\partial v} &= \frac{2v-1-(5v^2-6v+2)^{1/2}}{2(1-v)^2(5v^2-6v+2)^{1/2}} \neq 0, \quad 1/2 \leq v < 1. \end{aligned}$$

Thus, $u(v, t)$ is an admissible control, and in view of the equation $J(u(v, t), v) = 0$ the control is optimal, since $J(u(t), v) \geq 0$, $u(t) \in \Omega$.

2. Existence theorem

Let us give the conditions for the existence of a solution in the minimax control problem stated above.

Theorem

Let the right-hand side of the system (1) have the form $f(x, u, v) = C(x, v)u + d(x, v)$, where the components of the matrix $C(x, v)$ and of the vector $d(x, v)$ are continuous, along with their

partial derivatives with respect to x , in the set $X \times Q$. We assume that X , P , and Q are compact. In addition, P is a convex set. We assume that the integrand in (2) is continuous in $X \times P \times Q$, convex with respect to u in the set P for all fixed $(x, v) \in X \times Q$ and satisfies a Lipschitz condition with respect to $(x, u) \in X \times P$ with constant L for any fixed $v \in Q$.

If the set Q is not empty, then the expression

$$\min_{\Omega} \max_Q J(u(t), v).$$

reaches its minimum on some admissible control.

Proof. We shall first show that the set Q is weakly compact. We fix an arbitrary parameter of the set Q .

If we take the sequence of admissible controls $u_k(t)$, $k = 1, 2, \dots$, weakly convergent to the measurable function $\bar{u}(t)$, then it is sufficient to show that $\bar{u}(t)$ belongs to the set Q , since the set of all measurable controls with values in a convex compactum is weakly compact [5]. This means that $\bar{u}(t)$ takes values from the set P . The set of solutions of the system of differential equations $x_k(t)$ corresponding to the controls $u_k(t)$ is relatively compact (see the proof of our proposition). Hence we can assume that the sequence $x_k(t)$ is uniformly convergent to a function $\bar{x}(t)$. Using this fact, along with the weak convergence of the sequence $u_k(t)$ to $\bar{u}(t)$, and the linearity of the system of differential equations with respect to the control, we obtain

$$\begin{aligned} x_k(t) &= x^0 + \int_0^t [C(x_k, v)u_k + d(x_k, v)] dt \rightarrow x^0 \\ &+ \int_0^t [C(\bar{x}, v)\bar{u} + d(\bar{x}, v)] dt, \quad k \rightarrow \infty, \\ \lim_{k \rightarrow \infty} x_k(t) &= \bar{x}(t), \quad t \in [0, T]. \end{aligned}$$

The last equation implies that $\bar{x}(t)$ is the solution of the system of differential equations with the control $\bar{u}(t)$ and any fixed parameter v . Since the uniform convergence to $\bar{x}(t)$ holds in the interval $[0, T]$, and the set X is compact, then $\bar{x}(t)$ is defined in the same time interval and does not leave X .

For any fixed value of the parameter v , we shall show that the functional is weakly lower semi-continuous with respect to a control of Ω .

Let $u_k(t)$, $k = 1, 2, \dots$, be the sequence of admissible controls, and $x_k(t)$, $k = 1, 2, \dots$, the sequence of corresponding trajectories, weakly convergent to $\bar{u}(t)$ and uniformly convergent to the solution $\bar{x}(t)$ of the system with the control $\bar{u}(t)$, respectively. Let us show that the auxiliary functional

$$I(u(t)) = \int_0^T f(\bar{x}(t), u(t), v) dt$$

is weakly lower semi-continuous in the set of Lebesgue-measurable functions with values in P . Since the integrand is convex in P , we can easily see that the functional is convex with respect to the control, since $\bar{x}(t)$ and v are fixed. The auxiliary functional is continuous with respect to convergence of functions in $L_2 [0, T]$. This follows from the inequality

$$|I(\tilde{u}(t)) - I(u(t))| \leq \int_0^T |f^0(\bar{x}(t), \tilde{u}(t), v) - f^0(\bar{x}(t), u(t), v)| dt$$

$$\leq LT^{1/2} \left(\int_0^T |\tilde{u}(t) - u(t)|^2 dt \right)^{1/2}.$$

The weak lower semi-continuity of the auxiliary functional is obtained by applying the following theorem of functional analysis: a convex functional is weakly lower semi-continuous in a convex set of Banach space if and only if it is lower continuous in this set (see [6]).

In short, we have obtained the relation

$$\lim_{h \rightarrow \infty} I(u_h(t)) \geq I(\bar{u}(t)).$$

Using the following estimate for the difference between the values of the initial and the auxiliary functional for the chosen sequence of controls:

$$|J(u_h(t), v) - I(u_h(t))| \leq \int_0^T |f^0(x_h, u_h, v) - f^0(\bar{x}, u_h, v)| dt$$

$$\leq L \int_0^T |x_h - \bar{x}| dt \rightarrow 0, \quad k \rightarrow \infty,$$

we obtain the equation

$$\lim_{h \rightarrow \infty} J(u_h(t), v) = \lim_{h \rightarrow \infty} I(u_h(t)),$$

while from the definition of the auxiliary functional we obtain the equation $I(\bar{u}(t)) = J(\bar{u}(t))$.

On discarding the auxiliary functional, we get

$$\lim_{h \rightarrow \infty} J(u_h(t), v) \geq J(\bar{u}(t), v) \quad \forall v \in Q, \quad (4)$$

which implies that the functional (2) is weakly lower semi-continuous with respect to the control.

Given any admissible control, the functional has a maximum with respect to the parameter (see our proposition). We shall show that

$$\max_Q J(u(t), v) \quad (5)$$

is also a weakly lower semi-continuous functional in the set Ω . In fact, if the sequence of admissible controls $u_h(t)$, $h=1, 2, \dots$, is weakly convergent to $\bar{u}(t)$, then, by what has been proved, Eq. (4) holds. Since

$$\max_{v \in Q} J(u_h(t), v) \geq J(u_h(t), v),$$

we also have

$$\lim_{h \rightarrow \infty} \max_Q J(u_h(t), v) \geq \lim_{h \rightarrow \infty} J(u_h(t), v) \quad \forall v \in Q.$$

On combining these inequalities, we see that

$$\lim_{h \rightarrow \infty} \max_Q J(u_h(t), v) \geq \max_Q J(\bar{u}(t), v).$$

If, to the functional (5), defined in the weakly compact set Ω , we apply the theorem of functional analysis, to the effect that a weakly lower semi-continuous functional reaches its greatest lower bound in a weakly compact set, we complete the proof (see [7]).

Corollary. The theorem remains true if the set X is replaced by the space E^n , and the assumption that the set of admissible controls is not empty is replaced by the condition

$$\begin{aligned} \|C(x, v)\| &\leq c_1(1 + |x|), & \|d(x, v)\| &\leq c_2(1 + |x|) \\ \forall (x, v) &\in E^n \times Q, \end{aligned} \quad (6)$$

where $\|C(x, v)\|$ and $\|d(x, v)\|$ denote the norms of the matrix $C(x, v)$ and of the vector $d(x, v)$, and c_1, c_2 are positive constants.

Proof. Given any Lebesgue-measurable control with values in P , and any parameter of Q , the system of differential equations has a solution, defined in a certain segment (see [8]).

Satisfaction of the inequality (6) ensures that the solution can be continued into the interval $[0, T]$, so that the set Ω is the same as the set of Lebesgue-measurable functions in the interval $[0, T]$, with values in P .

Let us show that all the possible trajectories, corresponding to controls of Ω and parameters of Q , cannot leave a sphere of the space E^n . It can be assumed that the inequalities (6) hold in the Euclidean norm. We denote by $y(t)$ the Euclidean norm of the phase variable $x(t)$.

We then easily obtain the inequality

$$dy^2(t)/dt = 2(x(t), C(x(t), v)u(t) + d(x(t), v)) \leq 2c_0 y(1 + y),$$

where the constant

$$c_0 = c_1 \max_{u \in P} |u| + c_2,$$

and (g, f) denotes the scalar product of vectors g and f .

After integrating the differential inequality, we obtain an estimate, independent of the choice of admissible control and parameter: $y(t) \leq (1 + |x^0|)e^{c_0 t} - 1$. It remains to use the theorem.

3. Some generalizations

If, instead of a parameter, we use measurable vector functions $v(t)$ with values from Q , we can easily see that, if all the other conditions in the theorem are satisfied, we can assert that an admissible control exists, realizing

$$\min_{\Omega} \sup_{v(t)} J(u(t), v(t)).$$

If, at the same time, we consider a system of differential equations of the type $dx/dt = C(x)u(t) + D(x)v(t)$ and require in addition that the set Q be convex and the function $f^0(x, u, v)$ be concave with respect to $v \in Q$ for all fixed $(x, u) \in X \times P$, then we can prove in a similar way that there exist an admissible control $u(t)$ and a function $v(t)$, such that

$$\min_{u(t)} \max_{v(t)} J(u(t), v(t))$$

is reached on these functions.

This last result generalizes the result of [9], where a problem in a similar formulation was considered.

Translated by D. E. Brown

REFERENCES

1. GERMEIER, YU. B., *Introduction to operations research theory* (Vvedenie v teoriyu issledovaniya operatsii), Nauka, Moscow, 1971.
2. GABASOV, R., and KIRILLOVA, F. M., *The maximum principle in optimal control theory* (Printsip maksimuma v teorii optimal'nogo upravleniya), Nauka i Tekhnika, Minsk, 1974.
3. GURIN, L. G., Some topics in optimal control theory, *Diss. kand. fiz.-matem. n.*, VTs Akad. Nauk SSSR, Moscow, 1974.
4. VINOGRADOVA, T. K., and DEM'YANOV, V. F., On necessary conditions in minimax control problems, *Zh. vychisl. Mat. mat. Fiz.*, **14**, No. 1, 233-236, 1974.
5. LEE, E. B., and MARKUS, L., *Foundations of optimal control*, Wiley, 1967.
6. VASIL'EV, F. P., *Foundations of numerical methods for solving extremal problems (text of lectures)* (Osnovy chislennykh metodov resheniya ekstremal'nykh zadach (teksty lektsii)), No. 2, Izd-vo MGU, Moscow, 1973.
7. LYUSTERNIK, L. A., and SOBOLEV, V. I., *Elements of functional analysis* (Elementy funktsional'nogo analiza), Nauka, Moscow, 1965.
8. CODDINGTON, E. A., and LEVINSON, N., *Theory of ordinary differential equations*, McGraw, 1955.
9. PODINOVSKII, V. V., On the existence of a solution in minimax optimal process problems, *Izv. Akad. Nauk SSSR, Tekhn. kibernetika*, No. 1, 33-38, 1968.

A METHOD OF EVALUATING THE STATIONARY POINTS OF A GENERAL PROBLEM OF NON-LINEAR PROGRAMMING*

N. A. BOGOMOLOV and V. G. KARMANOV

Moscow

(Received 26 March 1976)

THE USE of the method of feasible directions (mfd) for finding the points of local minima of a non-convex function in a non-convex set is considered. The successive approximations are shown to be convergent to the set of stationary points, and in particular, to the set of points at which the necessary conditions for a local minimum are satisfied.

To find the points of local minima in a general problem of non-linear programming, the only realistic approach is to use a relaxation method in which, when finding the direction of descent from a point x , account is taken solely of the local properties of the function requiring minimization, and the local properties of the set in which this function is defined. Of the available determinate methods, the mfd satisfies these conditions [1-4].

Consider the problem of finding the minima of the differentiable function $\varphi(x)$ in a closed set X of n -dimensional Euclidean space.

*Zh. vychisl. Mat. mat. Fiz., 17, 72-78, 1977.

Let $X = \{x | f_i(x) \geq 0, i=1, 2, \dots, m\}$, $f_i(x)$ be given differentiable functions. The mfd consists in constructing a sequence of points $\{x_k\}$ from the expression $x_{k+1} = x_k - \beta_k s_k$. Here, $x_k \in X$ is the point evaluated at the previous iteration, and $-s_k$ is a feasible direction at the point x_k , i.e., a direction such that small displacements along it from the point x_k do not go outside the set X ; and finally, let β_k define the step length. Here, s_k and β_k are chosen in such a way that

$$\varphi(x_{k+1}) \leq \varphi(x_k), \quad k=0, 1, \dots$$

We shall consider the auxiliary problem of finding the number σ and the vector s such that

$$\begin{aligned} \sigma &\rightarrow \max, \\ \langle f'_i(x), s \rangle + \sigma &\leq 0, \quad i \in I, \quad -\langle \psi'(x), s \rangle + \sigma \leq 0, \\ \langle s, s \rangle &\leq 1. \end{aligned} \quad (1)$$

Here, $\langle \psi'(x), s \rangle$ is the scalar product of the gradient of the function $\psi(x)$ and the vector s .

Let $\bar{\sigma}(x, \varepsilon)$ and $\bar{s}(x, \varepsilon)$ denote the solutions of problem (1) for $I = I(x, \varepsilon) = \{i : 0 \leq f_i(x) \leq \varepsilon\}$, let ε be a positive number, and let $\bar{\sigma}(x, 0)$ and $\bar{s}(x, 0)$ be the solutions of problem (1) for $I = I(x, 0) = \{i : f_i(x) = 0\}$. In order for the direction $-s$, $\|s\|=1$, to be feasible at the point $x \in X$ it is sufficient that $\sigma > 0$ and s satisfy the inequalities $\langle f'_i(x), s \rangle + \sigma \leq 0$ for all $i \in I(x, \varepsilon)$ for at least one $\varepsilon \geq 0$ (see e.g., [4]). Let the direction $-s$ be feasible at the point $x \in X$. We define the distance ζ from the point x to the nearest boundary point of the set X along the direction $-s$. Since the direction $-s$ is feasible at the point x , a number $\bar{\beta} > 0$ exists such that the point $x - \beta s \in X$ for all $\beta \in [0, \bar{\beta}]$. The quantity $\zeta = \sup \bar{\beta}$ (if it is finite) denotes the length of the maximum interval $[x, x - \zeta s]$, belonging entirely to the set X . Here, $y = x - \zeta s$ is the boundary point of the set X . If $\zeta = +\infty$, then the ray $x - \beta s$, $\beta \geq 0$, belongs to the set X . If $x = x_k$ and $s = s_k$ we shall write $\zeta = \zeta_k$.

Scheme of the method. As the initial approximation x_0 we can choose any element of the set X , while ε_0 is chosen from the semi-interval $(0, 1]$. Assume that x_k and ε_k have been evaluated as a result of the k -th iteration. Let us describe the $(k+1)$ -th iteration.

Step A. On solving problem (1) for $I = I(x_k, \varepsilon_k)$, we can find the admissible σ_k and s_k , $\|s_k\|=1$, such that $\sigma_k \geq \bar{\sigma}_k(x_k, \varepsilon_k)$, where $0 < \bar{\sigma}_k \leq \bar{\sigma}_k \leq 1$.

Step B. If $\sigma_k \geq \varepsilon_k$, we evaluate β_k . Usually, the β_k are evaluated by solving a problem of one-dimensional minimization. Then, β_k has to satisfy the conditions

$$\begin{aligned} \varphi(x_k - \beta_k s_k) &\leq (1 - \lambda_k) \varphi(x_k) + \lambda_k \omega_k, \quad 0 < \lambda_k \leq \lambda_k \leq 1, \\ \omega_k &= \inf_{0 \leq \beta \leq \zeta_k} \varphi(x_k - \beta s_k). \end{aligned} \quad (2)$$

The numbers β_k can also be chosen as follows. Let $\bar{\beta}_k$ be the maximum of the numbers which satisfy the relations

$$\varphi(x_k) - \varphi(x_k - \beta_k s_k) \geq \frac{1}{2} \beta_k s_k, \quad 0 \leq \beta_k \leq \zeta_k. \quad (3)$$

As β_k we can take any number which satisfies the inequalities (3) and the condition $\beta_k \geq \alpha \beta_k$ for any $\alpha \in (0, 1]$.

Finally, we evaluate $x_{k+1} = x_k - \beta_k s_k$, put $\varepsilon_{k+1} = \varepsilon_k$ and pass to step A.

Step C. If $0 < \sigma_k < \varepsilon_k$, we put $x_{k+1} = x_k$, $\varepsilon_{k+1} = \gamma_k \varepsilon_k$, where $0 < \gamma_k \leq \gamma < 1$, and we pass to step A. If $\sigma_k = 0$, we evaluate $\tilde{\sigma}(x_k, 0)$, by solving problem (1) for $I = I(x_k, 0)$. If $\tilde{\sigma}(x_k, 0) = 0$, the process is terminated. Otherwise, we put $x_{k+1} = x_k$, $\varepsilon_{k+1} = \gamma_k \varepsilon_k$, where $0 < \gamma_k \leq \gamma < 1$, and pass to step A.

The convergence of our method will be proved under the following assumptions.

Condition 1. The functions $\varphi(x)$ and $f_i(x)$, $i=1, 2, \dots, m$, belong to the class $C^{1,1}(X)$.

Condition 2. A number $M > 0$ exists, such that $\|f'_i(x)\| \leq M$ for all $x \in X$, $i=1, 2, \dots, m$.

Condition 3. The set $X^* = \{x^* \in X \mid \tilde{\sigma}(x^*, 0) = 0\}$ is not empty.

Condition 4. $\inf_{x \in X} \varphi(x) > -\infty$.

Condition 5. The sequence $\{x_k\}$ is compact.

Let us explain condition 3. If the function $\varphi(x)$ and the set X are convex and satisfy Slater's condition, then the condition $\tilde{\sigma}(x^*, 0) = 0$ is necessary and sufficient for $\varphi(x)$ to have a global minimum at the point x^* . If only the set X is convex, then X^* is the set of points at which the necessary conditions for a local minimum are satisfied. In the general case, to these points may also be added a series of others, e.g., the points at which no feasible directions exist. In the present paper the convergence of the sequence $\{x_k\}$ to the set of stationary points X^* is investigated, so that condition 3 is natural.

Notice that conditions (2) and (3) ensure that the sequence $\{\varphi(x_k)\}$ is not monotonically increasing.

The convergence will be proved under the assumption that the sequence $\{x_k\}$ is finite, since otherwise, in accordance with step C, $\tilde{\sigma}(x_k, 0) = 0$, i.e. $x_k \in X^*$.

Lemma 1

For all σ and s , satisfying the conditions

$$\langle f'_i(x), s \rangle + \sigma \leq 0, \quad i \in I(x, \varepsilon), \quad x \in X, \quad \langle s, s \rangle \leq 1, \quad \sigma \geq \varepsilon \geq 0, \quad (4)$$

we have

$$\xi \geq C\varepsilon, \quad (5)$$

where $C = \min \{1/M, 1/L\}$.

If $\psi(x) \in C^{1,1}(X)$, then a number $L > 0$ exists such that, given any interval $[x, y]$, belonging entirely to the set X , we have $\|\psi'(x) - \psi'(y)\| \leq L\|x - y\|$. Since the number of functions $\varphi(x), f_i(x)$ is finite, a Lipschitz constant common to all the functions will exist.

Proof. Obviously, it is sufficient to take the case $\zeta < +\infty$. Since the point $y = x - \zeta s$ is on the boundary (by definition of ζ), a number i will exist such that $f_i(y) = 0$. If $f_i(x) > \varepsilon$ at the point x , then $\varepsilon < f_i(x) = |f_i(x) - f_i(y)| \leq M\|x - y\| = M\zeta$, whence

$$\zeta > \varepsilon/M. \quad (6)$$

Assume that $0 \leq f_i(x) \leq \varepsilon$, i.e. $i \in I(x, \varepsilon)$ at the point x . Put $\psi_i(\beta) = f_i(x - \beta s)$ and notice that $\psi_i(\beta) \geq 0$ for $\beta \in [0, \zeta]$ and $\psi_i(\zeta) = 0$. It can easily be seen that $[d\psi_i(\beta)/d\beta]|_{\beta=\zeta} \leq 0$, i.e., $\langle f'_i(y), s \rangle \geq 0$. From the condition $i \in I(x, \varepsilon)$ and the condition that σ and s satisfy system (4), we have $\langle f'_i(x), s \rangle \leq -\sigma$, so that

$$\begin{aligned} \varepsilon &\leq \sigma \leq -\langle f'_i(x), s \rangle \leq \langle f'_i(y), s \rangle - \langle f'_i(x), s \rangle \\ &\leq \|f'_i(y) - f'_i(x)\| \|s\| \leq L\|y - x\| = L\zeta, \end{aligned}$$

whence

$$\zeta \geq \varepsilon/L. \quad (7)$$

From (6) and (7) we obtain $\zeta \geq \varepsilon \min \{1/M, 1/L\}$, i.e., the inequality (5).

Lemma 2

If the point x_{k+1} is constructed in accordance with the scheme of the method, and $\sigma_k \geq \varepsilon_k \geq 0$, we have

$$\varphi(x_k) - \varphi(x_{k+1}) \geq \frac{1}{2} C \alpha \lambda \varepsilon_k^2 \quad (8)$$

for fixed $\alpha \in (0, 1]$ and $\lambda \in (0, 1]$.

The proof follows from the inequality*

$$\varphi(x_k) - \varphi(x_{k+1}) \geq \frac{1}{2} \alpha \lambda \sigma_k \min \left\{ \zeta_k, \frac{\sigma_k}{L} \right\} \quad (9)$$

and inequality (5).

Lemma 3

For any point $\bar{x} \in X$ numbers $\bar{\varepsilon} = \bar{\varepsilon}(\bar{x}) > 0$ and $\bar{\delta} = \bar{\delta}(\bar{x}) > 0$ exist such that, for all $x \in U_{\bar{\delta}}(\bar{x}) = \{x \in X : \|x - \bar{x}\| \leq \bar{\delta}\}$ we have $I(x, \varepsilon) \subset I(\bar{x}, 0)$.

Proof. Put $J = \{i = 1, 2, \dots, m\}$ and let us choose

$$\bar{\varepsilon} = \frac{1}{2} \min_{i \in J \setminus I(\bar{x}, 0)} \{f_i(\bar{x})\}.$$

*See [4], p.239.

Notice that $\bar{\epsilon} > 0$, since, for $i \in J \setminus I(\bar{x}, 0)$ we have $f_i(\bar{x}) > 0$.

Since the functions $f_i(x)$ are continuous and the number of numbers $i, i \leq m$, is finite, a number $\delta(\bar{\epsilon}) > 0$ will exist such that, for all $x \in U_\delta(\bar{x})$ and all $i \in J$ we have $|f_i(\bar{x}) - f_i(x)| \leq \bar{\epsilon}$. Since $f_i(\bar{x}) \geq 2\bar{\epsilon}$, $i \in J \setminus I(\bar{x}, 0)$, then $f_i(x) \geq \bar{\epsilon}$ for all $x \in U_\delta(\bar{x})$ and all $i \in J \setminus I(\bar{x}, 0)$. Hence, for $\epsilon \in [0, \bar{\epsilon}]$ for any $i \in J \setminus I(\bar{x}, 0)$ we have $i \in J \setminus I(x, \epsilon)$, and hence we get the inclusion $I(x, \epsilon) \subset I(\bar{x}, 0)$.

Let $\bar{\sigma} = \bar{\sigma}(\bar{x}, 0)$ and $\bar{s} = \bar{s}(\bar{x}, 0)$ be the solutions of the problem (1) for $x = \bar{x}$ and $I = I(\bar{x}, 0)$.

Lemma 4

If $I(x, \epsilon) \subset I(\bar{x}, 0)$ and $\bar{\sigma} > 0$, then a number $\delta = \delta(\bar{\sigma}) > 0$, exists, such that, for all $x \in U_\delta(\bar{x})$ we have $\bar{\sigma}(x, \epsilon) \geq \bar{\sigma}/2$.

Proof. Since $\varphi'(x)$ and $f'_i(x)$, $i = 1, 2, \dots, m$, are continuous, a number $\delta(\bar{\sigma}) > 0$, will exist, such that, for all $x \in U_\delta(\bar{x})$ we have $\|\varphi'(x) - \varphi'(\bar{x})\| \leq \bar{\sigma}/2$ and $\|f'_i(x) - f'_i(\bar{x})\| \leq \bar{\sigma}/2$, $i \in I(\bar{x}, 0)$. Given any $x \in U_\delta(\bar{x})$ and any $i \in I(x, \epsilon) \subset I(\bar{x}, 0)$

$$\begin{aligned} 0 &\geq \langle f'_i(\bar{x}), \bar{s} \rangle + \bar{\sigma} = \langle f'_i(x), \bar{s} \rangle + \bar{\sigma} + \langle f'_i(\bar{x}) - f'_i(x), \bar{s} \rangle \\ &\geq \langle f'_i(x), \bar{s} \rangle + \bar{\sigma} - \|f'_i(\bar{x}) - f'_i(x)\| \|\bar{s}\| \geq \langle f'_i(x), \bar{s} \rangle + \bar{\sigma}/2, \end{aligned}$$

and similarly,

$$0 \geq -\langle \varphi'(x), \bar{s} \rangle + \bar{\sigma}/2.$$

Hence $\sigma = \bar{\sigma}/2$ and $s = \bar{s}$ will satisfy the conditions of problem (1) with $I = I(x, \epsilon)$ for any $x \in U_\delta(\bar{x})$. But $\bar{\sigma}(x, \epsilon)$ and $\bar{s}(x, \epsilon)$ represent the solution of the problem of maximizing σ under the same conditions, so that $\bar{\sigma}(x, \epsilon) \geq \bar{\sigma}/2$.

Theorem

If conditions 1-5 are satisfied, then

$$\lim_{k \rightarrow \infty} \rho(x_k, X^*) = \lim_{k \rightarrow \infty} \inf_{x^* \in X^*} \|x_k - x^*\| = 0.$$

Proof. Let K be the collection of all indices of the sequence $\{x_k\} : K = \{k=0, 1, \dots\}$. Assume that $\epsilon_k \geq \epsilon > 0$ for some $\epsilon > 0$ and for all $k \geq k_0$. Then, in accordance with the scheme of our method, a number $k_0 \in K$, will exist, such that $\sigma_k \geq \epsilon_k \geq \epsilon > 0$ for all $k \geq k_0$. From (5) we have $\xi_k \geq C\epsilon_k \geq C\epsilon$, and then, in view of (8), we obtain $\varphi(x_k) - \varphi(x_{k+1}) \geq C\alpha\lambda\epsilon^2/2$ for all $k \geq k_0$. But the sequence $\{\varphi(x_k)\}$ is convergent (since it is monotonic and bounded), which contradicts the previous inequality. In short, $\epsilon_k \rightarrow 0$, $k \rightarrow \infty$. In accordance with the scheme of the method, a collection of indices $K_1 \subset K$ exists, such that $\sigma_k \rightarrow 0$, $k \in K_1$, $k \rightarrow \infty$.

We shall now show that all the limit points of the sequence $\{x_k\}$ belong to the set X^* .

We consider two cases.

Case 1. Let \bar{x} be the unique limit point, i.e.,

$$\lim_{k \rightarrow \infty} x_k = \bar{x} \in X.$$

Assume that $\bar{x} \in X \setminus X^*$. Then, $\bar{\sigma}(\bar{x}, 0) = \bar{\sigma} > 0$, and hence, by Lemma 3, numbers $\varepsilon > 0$ and $\delta(\varepsilon) > 0$, exist such that $I(x, \varepsilon) \subset I(\bar{x}, 0)$ for all $\varepsilon \in [0, \varepsilon]$ and $x \in U_\delta(\bar{x})$. Since $\varepsilon_k \rightarrow 0$ and $x_k \rightarrow \bar{x}$, $k \rightarrow \infty$, a number k_0 will exist such that $x_k \in U_\delta(\bar{x})$ and $\varepsilon_k < \varepsilon$ for all $k \geq k_0$, and hence $I(x_k, \varepsilon_k) \subset I(\bar{x}, 0)$. But then, by Lemma 4, the inequality $\bar{\sigma}(x_k, \varepsilon_k) \geq \bar{\sigma}/2$, becomes valid for $k \geq k_0$, which contradicts the fact that $\sigma_k \rightarrow 0$, $k \in K_1$, $k \rightarrow \infty$.

Case 2. Assume that there is a limit point $\tilde{x} \in X$ of the sequence $\{x_k\}$, which differs from \bar{x} : $\tilde{x} \neq \bar{x}$. As before, we shall assume that $\bar{x} \in X \setminus X^*$. Then $\delta > 0$ exists such that $\tilde{x} \in X \setminus U_\delta(\bar{x})$. Since \bar{x} and \tilde{x} are two distinct limit points of the sequence, there will exist, for any integer N , a number $k \geq N$ and a number $m \geq 1$, such that $x_k \in U_{\delta/2}(\tilde{x})$, $x_{k+i} \in U_\delta(\tilde{x})$, $i = 0, 1, \dots, m-1$, $x_{k+m} \in X \setminus U_\delta(\tilde{x})$.

We now use the inequality (9):

$$\begin{aligned} \varphi(x_k) - \varphi(x_{k+m}) &= \sum_{i=k}^{k+m-1} \varphi(x_i) - \varphi(x_{i+1}) \\ &\geq \frac{1}{2} \alpha \lambda \sum_{i=k}^{k+m-1} \sigma_i \min \left\{ \zeta_i, \frac{\sigma_i}{L} \right\}, \end{aligned}$$

and since $x_i \in U_\delta(\tilde{x})$, $i = k, k+1, \dots, k+m-1$, we have $\sigma_i \geq \bar{\sigma}/2$, $i = k, k+1, \dots, k+m-1$. In addition, $\zeta_i \geq \beta_i$ for all numbers i , so that

$$\varphi(x_k) - \varphi(x_{k+m}) \geq \frac{1}{4} \alpha \lambda \bar{\sigma} \sum_{i=k}^{k+m-1} \min \left\{ \beta_i, \frac{\bar{\sigma}}{2L} \right\}.$$

Notice that

$$\Sigma = \sum_{i=k}^{k+m-1} \min \left\{ \beta_i, \frac{\bar{\sigma}}{2L} \right\} \geq \min \left\{ \sum_{i=k}^{k+m-1} \beta_i, \frac{\bar{\sigma}}{2L} \right\}.$$

For, if all the $\beta_i \leq \bar{\sigma}/2L$, we have

$$\Sigma = \sum_{i=k}^{k+m-1} \beta_i.$$

If, for at least one j , we have $\beta_j > \bar{\sigma}/2L$, then $\Sigma > \bar{\sigma}/2L$. Since $\|x_{k+m} - x_k\| \geq \delta/2$, we have

$$\sum_{i=k}^{k+m-1} \beta_i \geq \frac{\delta}{2},$$

and hence

$$\varphi(x_k) - \varphi(x_{k+m}) \geq \frac{1}{4} \alpha \lambda \bar{\sigma} \min \left\{ \frac{\delta}{2}, \frac{\bar{\sigma}}{2L} \right\} = \text{const} > 0.$$

This last inequality contradicts the convergence of the sequence $\{\varphi(x_k)\}$.

In short, any limit point of the sequence $\{x_k\}$ belongs to the set X^* . Let us show that

$$\lim_{k \rightarrow \infty} \rho(x_k, X^*) = 0.$$

If this is not the case, then a number $\Delta > 0$ and a subsequence $\{x_k\}$, $k \in K_2 \subset K$, will exist, such that $\rho(x_k, X^*) > \Delta$ for all $k \in K_2$. But in view of the compactness and what has been said above, there will be a collection of numbers $K_3 \subset K_2$ for which we have

$$\lim_{k \rightarrow \infty, k \in K_2} x_k = \bar{x} \in X^*,$$

which contradicts our assumption that $\rho(x_k, X^*) > \Delta$ for all $k \in K_2$.

Notice that the convergence theorem is proved for any initial approximation $x_0 \in X$. In actual problems we usually have information about a point x_0 which is reasonably close to the required point $x^* \in X^*$,

Let the point x_0 belong to some connectivity component set X_0 of the set $\{x \in X \mid \varphi(x) \leq \varphi(x_0)\}$. If the set X_0 is bounded, and for any $x^* \in X^* \cap X_0$ we have

$$\varphi(x^*) = \min_{x \in X_0} \varphi(x),$$

then

$$\lim_{k \rightarrow \infty} \varphi(x_k) = \varphi(x^*), \quad \lim_{k \rightarrow \infty} \rho(x_k, X^* \cap X_0) = 0.$$

For, since $\varphi(x_{k+1}) \leq \varphi(x_k)$ for all k , then $x_k \in X_0$, $k=0, 1, \dots$. By the convergence theorem, a collection of indices $K_1 \subset K$ exists, such that $x_k \rightarrow \bar{x}$, $k \in K_1$, $k \rightarrow \infty$, and $\delta(\bar{x}, 0) = 0$, so that $\bar{x} \in X^* \cap X_0$. In view of the convergence of the sequence $\{\varphi(x_k)\}$ we get

$$\lim_{k \rightarrow \infty} \varphi(x_k) = \varphi(\bar{x}).$$

In conclusion it may be mentioned that our entire discussion also holds for the case when the condition $\langle s, s \rangle \leq 1$ in problem (1) is replaced by the condition

$$\max_{j=1, 2, \dots, n} |s_j| \leq 1;$$

here, s_j denotes the j -th component of the vector $s = (s_1, \dots, s_n)$. In this case, problem (1) becomes a problem of linear programming.

Translated by D. E. Brown

REFERENCES

1. ZOUTENDIJK, G., *Methods of feasible directions*, Elsevier, 1960.
2. PSHENICHNYI, B. N., and DANILIN, YU. M., *Numerical methods in extremal problems* (Chislennye metody v ekstremal'nykh zadachakh), Nauka, Moscow, 1975.
3. POLAK, E., *Numerical methods of optimization* (Chislennye metody optimizatsii), Mir, Moscow, 1974.
4. KARMANOV, V. G., *Mathematical programming* (Matematicheskoe programmirovaniye), Nauka, Moscow, 1975.

REGULARITY CONDITIONS AND NECESSARY CONDITIONS FOR A MAXIMIN WITH CONNECTED VARIABLES*

V. V. FEDOROV

Moscow

(Received 16 June 1975)

REGULARITY conditions are introduced for non-convex problems. In conjunction with the method of penalties, they enable the necessary conditions for an optimum to be obtained in minimax problems.

Introduction

The directional differentiability of a minimum function was proved in [1, 2], thereby enabling the necessary conditions for optimality to be derived in maximin problems with connected variables. Since, however, the assumptions ensuring the existence of directional derivatives in the case of connected variables are quite rigid, it seems useful to obtain the necessary conditions for optimality under wider assumptions. This can be done using the method of penalty functions, which has only quite recently come to be systematically used to derive optimality conditions (see e.g., [3, 4]). The method of penalties proves to be especially effective in this sense in complex maximin problems, thanks to the wide range of convergence theorems that is now available for it [4].

In general terms, the method of obtaining the necessary conditions is as follows. The initial problem is first reduced to a parametric family of simpler problems which have previously been investigated. We then pass to the limit with respect to the penalty parameter in the optimality conditions for the penalty problems, and thereby obtain the optimality conditions in the initial problem.

An approach of this kind offers a basis for considering the method of penalty functions as an "algorithm" for stating the optimality conditions in extremal problems. As distinct from the existing general schemes for analyzing extremal problems [5, 6], the method of penalties in minimax problems does not lead to necessary conditions for optimality of a general type (of the Euler-Lagrange equation type). Each new more complicated problem has to be analyzed on the basis of results previously obtained, while utilizing the "algorithm" stated above. It should be mentioned that, as a rule, each such "step" does not involve too serious difficulties, provided that we arrange for the problems to become only gradually more complicated.

In our view, there is no justification for attempting to obtain optimality conditions of the most general possible kind in minimax problems. First, such conditions (even if they could be obtained) would be extremely complicated and unwieldy. Second, the attempt would require as a preliminary a refined idea of what is meant by a general minimax problem. The latter is unattainable in principle. The point is that any minimax problem can be regarded as the consequence of applying the principle of the best guaranteed result (or other optimality principle) in some game

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 79-90, 1977.

in certain strategy sets. Since new systems and methods of operation (strategies) are constantly arising, new minimax problems will make their appearance. It follows from what has been said that the above-mentioned feature of the general scheme of obtaining necessary conditions on the basis of the method of penalties can be looked on as a merit rather than a drawback. Of course, this does not mean that concrete classes of minimax problems cannot be more carefully studied with the aid of traditional schemes.

In the present paper we show that, if discontinuous penalties are used [4], the well-known conditions for the exact solution of a problem with constraints by the penalty function method are also conditions for regularity of the problem, and lead to stronger conditions for optimality. This is also true for the maximin with connected variables. There is no difficulty in extending all our conclusions to the problem of seeking a multiple maximin with constraints, so that a wide range of problems in operations research and the theory of games is covered.

1. Regularity conditions

We consider the problem of finding

$$\max_{x \in A} F(x) = F(x^0), \quad (1)$$

where

$$A = \{x \in X \mid \varphi_i(x) \geq 0, \quad 1 \leq i \leq m\}. \quad (2)$$

Definition 1. We shall say that the functional constraints, $\varphi_i(x) \geq 0$, $1 \leq i \leq m$, specifying the set A , are regular, if numbers K and $\delta > 0$ exist, such that, for all $x \in (V_\delta(A) \setminus A) \cap X$ we have

$$\min_{1 \leq i \leq m} \varphi_i(x) \leq -K \rho(x, A). \quad (3)$$

Here and below, ρ is the metric in X , and $V_\delta(A)$ is the δ -neighbourhood of the set A .

Whether the regularity conditions (3) are satisfied will naturally depend on how the set A is described, i.e., on the form of the functions $\varphi_i(x)$. However, any set A can formally be specified in the form (2) by regular constraints. For this, we have to put $m = 1$ and $\varphi_1(x) = -\rho(x, A)$.

For conditions (3) to be satisfied, it is sufficient that at least one of the following conditions be satisfied:

- 1) $\varphi_i(x)$ are linear functions, and X is a convex set of Euclidean space E_n [4];
- 2) $\varphi_i(x)$ are concave and satisfy Slater's condition in a bounded convex set X , i.e., a point $\tilde{x} \in X$, exists, for which $\varphi_i(\tilde{x}) > 0$, $1 \leq i \leq m$ [4];
- 3) X is a finite set.

In short, the inequality (3) expresses a characteristic property inherent in the well-known regularity conditions [7] in convex programming (at least for bounded X). At the same time, condition (3) holds for a class of functions wider than the class of concave functions. Let us mention an example:

$$A = \{x \in X \mid \max_{1 \leq i \leq m} \varphi_i(x) \geq 0\},$$

where $\varphi_i(x)$ are concave functions on the bounded convex set X , and points $\tilde{x}_i \in X$ exist, such that $\varphi_i(\tilde{x}_i) > 0$, $1 \leq i \leq m$.

A similar example may be constructed with linear $\varphi_i(x)$.

Condition (3) implies geometrically that, in a δ -neighbourhood of the set A , the function

$$\varphi(x) = \min_{1 \leq i \leq m} \varphi_i(x)$$

decreases at least as fast as a linear function of the distance to the set A . While it may not be easy to check condition (3) directly, the stock of such functions is clearly considerable.

The following theorem, proved in [4], will be required below:

Theorem 1

Let the functions $\varphi_i(x)$ be continuous, let $F(x)$ satisfy a Lipschitz condition in the compact set X , and let the regularity conditions (3) hold. Then, for all sufficiently large C ,

$$\begin{aligned} 0 \leq \max_{x \in X} L_q(x, C) - \max_{x \in A} F(x) &\leq \begin{cases} BC^{-1/(q-1)}, & q > 1, \\ 0, & q \leq 1, \end{cases} \\ \rho(x_q^0(C), A) &\leq \begin{cases} BC^{-1/(q-1)}, & q > 1, \\ 0, & q \leq 1, \end{cases} \end{aligned}$$

where

$$\begin{aligned} L_q(x, C) &= F(x) - C \sum_{i=1}^m |\min(0, \varphi_i(x))|^q, \\ L_q(x_q^0(C), C) &= \max_{x \in X} L_q(x, C), \quad x_q^0(C) \in X, \end{aligned}$$

and B is a constant independent of C .

Theorem 1 establishes an error estimate in the case when problem (1), (2) is solved by the method of penalties. In particular, the penalty function $L_1(x, C)$ gives the exact solution of the problem for a certain finite C . We know [4, 8] that, for a concave problem (1), (2), this is equivalent to the Lagrange function having a saddle point.

We shall now obtain the necessary conditions, satisfied by any solution of problem (1), (2).

Theorem 2

In (1), (2), let the functions $F(x)$ and $\varphi_i(x)$ be continuously differentiable in the convex closed set $X \subseteq E_n$. Then, numbers $\lambda_0, \lambda_1, \dots, \lambda_m \geq 0$ which do not vanish simultaneously, exist, such that

$$\begin{aligned} -\left\{ \lambda_0 F'(x^0) + \sum_{i=1}^m \lambda_i \varphi_i'(x^0) \right\} &\in K_{X^*}(x^0), \\ \lambda_i \varphi_i(x^0) &= 0, \quad 1 \leq i \leq m. \end{aligned} \tag{4}$$

If, in addition, the regularity conditions (3) hold, then $\lambda_0 > 0$ (it can be assumed without loss of generality that $\lambda_0 = 1$).

Here, $K^*_X(x^0)$ is the cone conjugate to the cone of feasible directions of the set X at the point x^0 .

Proof. 1. We put $F_1(x) = F(x) - \|x - x^0\|^2$. Then $F_1(x)$ has a unique realization x^0 of its maximum in X . All our remaining arguments will be carried out in the compact set $S \cap X$, where S is the closed sphere, center x^0 . Since all the conditions for convergence of the method of penalties hold in $S \cap X$, we have

$$\max_{x \in A} F(x) = \lim_{C \rightarrow \infty} \max_{x \in S \cap X} \left\{ L_2(x, C) = F_1(x) - \sum_{i=1}^m C [\min(0, \varphi_i(x))]^2 \right\}$$

and the realizations $x^0(C)$ of the maximum $L_2(x, C)$ in $S \cap X$ are convergent to x^0 as $C \rightarrow \infty$. For sufficiently large C , we write down the condition for an extremum of $L_2(x, C)$ in $S \cap X$:

$$-\left\{ F'_1(x^0(C)) - 2C \sum_{i=1}^m \min(0, \varphi_i(x^0(C))) \varphi'_i(x^0(C)) \right\} \in K^*_X(x^0(C)).$$

We now normalize (5) to

$$m(C) = 1 + \sum_{i=1}^m \mu_i(C), \text{ where } \mu_i(C) = -2C \min(0, \varphi_i(x^0(C))),$$

i.e., we introduce $\lambda_0(C) = 1/m(C)$, $\lambda_i(C) = \mu_i(C)/m(C)$ and pass to the limit in (5) as $C \rightarrow \infty$. Then, a sequence $\{C_k\} \rightarrow \infty$ can be chosen, such that $\lambda_0(C_k) \rightarrow \lambda_0$, $\lambda_i(C_k) \rightarrow \lambda_i$ (since

$$\sum_{i=0}^m \lambda_i(C) = 1, \quad \lambda_i(C) \geq 0) \text{ and at the same time, } x^0(C_k) \rightarrow x^0.$$

The mapping $x \rightarrow K^*_X(x)$ is closed, and $F'_1(x)$, $\varphi'_i(x)$ are continuous; hence (4) follows from (5). Further, if $\varphi_i(x^0) > 0$, then, for sufficiently large C_k , the coefficient $\mu_i(C_k) = 0$, i.e., the conditions of supplementary non-rigidity $\lambda_i \varphi_i(x^0) = 0$ are satisfied.

2. Now let the regularity conditions (3) hold. We shall show that the sum $\sum_{i=1}^m \mu_i(C)$ is then uniformly bounded with respect to C . This is all that is needed to complete the proof of the theorem, since, on passing to the limit in relation (5) as $C_k \rightarrow \infty$, the coefficient of $F'(x^0)$ will be equal to 1.

We have

$$\sum_{i=1}^m \mu_i(C) \leq m |2C \min_{1 \leq i \leq m} \min(0, \varphi_i(x^0(C)))| \leq m 2CN \rho(x^0(C), A),$$

where N is the Lipschitz constant of the functions $\varphi_i(x)$ in the sphere S . By Theorem 1,

$$\rho(x^0(C), A) \leq O(1/C); \text{ hence } \sum_{i=1}^m \mu_i(C) \text{ is bounded. The theorem is proved.}$$

Notes. 1. The scheme of proof of our theorem realizes the "algorithm" mentioned in the Introduction, for obtaining necessary conditions for optimality in problem (1), (2). In effect, the

problem has been reduced by the method of penalties to unconstrained optimization of the function $L_2(x, C)$, thereby enabling the necessary conditions for an unconstrained extremum to be applied. For problems more complicated than (1), (2), different theorems on the convergence of the method of penalty functions have to be employed, along with the relevant necessary conditions (see e.g., Sections 2 and 3, and also [3, 4]).

2. The theorem has the greatest interest in the case when $\lambda_0 = 1$, since in other cases the necessary conditions (4) are in no way connected with the function to be optimized $F(x)$.

3. As we indicated above, the regularity conditions are not merely satisfied in convex problems. Hence Theorem 2 establishes the existence of Lagrange factors $\lambda_0=1, \lambda_1, \dots, \lambda_m \geq 0$ for a wider class of problems.

4. When the regularity conditions (3) hold, the theorem can be proved in a different way. In fact, by Theorem 1, problem (1), (2) reduces for finite C_0 to seeking the maximin

$$\begin{aligned} \max_{x \in A} F(x) &= \max_{x \in S \cap X} \left\{ F(x) + C_0 \sum_{i=1}^m \min(0, \varphi_i(x)) \right\} \\ &= \max_{x \in S \cap X} \min_{0 \leq \lambda_i \leq C_0, 1 \leq i \leq m} \left\{ F(x) + \sum_{i=1}^m \lambda_i \varphi_i(x) \right\}. \end{aligned}$$

It remains only to apply the necessary conditions for a maximin [2-4].

2. Maximin with connected variables

Let us now turn to the problem of seeking

$$\sup_{x \in X} \min_{y \in B(x)} F(x, y)$$

and the point $x^0 \in X$ (if it exists), realizing

$$\min_{y \in B(x^0)} F(x^0, y) = \sup_{x \in X} \min_{y \in B(x)} F(x, y). \quad (6)$$

The many-valued mapping $B(x)$ can be assumed to be specified in the form

$$B(x) = \{y \in Y \mid g_j(x, y) \geq 0, 1 \leq j \leq m\}$$

and to be non-empty for all $x \in X$.

For the existence of an optimal strategy x^0 , it is sufficient that the mapping $B(x)$ be continuous in the Hausdorff metric, and $F(x, y)$ continuous in the compact sets X, Y . We know [4] that the sufficient condition for $B(x)$ to be Hausdorff-continuous is

$$\overline{B^0}(x) = B(x) \quad \forall x \in X,$$

where $B^0(x) = \{y \in Y \mid g_j(x, y) > 0, 1 \leq j \leq m\}$, and \overline{B} denotes the closure of the set B .

We shall use below a different sufficient condition for $B(x)$ to be continuous, following from the regularity conditions.

Definition 2. The mapping $B(x)$ is regular at the point $x^0 \in X$, if numbers $K, \delta > 0$ exist, such that, for all $x \in V_\delta(x^0)$ and all $y \in (V_\delta(B(x)) \setminus B(x)) \cap Y$ we have

$$\min_{1 \leq j \leq m} g_j(x, y) \leq -K \rho(y, B(x)). \quad (7)$$

The mapping $B(x)$ is regular in the set X if it is regular at any point $x \in X$ with fixed parameters $K, \delta > 0$.

If $B(x)$ is a constant mapping, the definition 2 is obviously the same as the definition 1.

It is easy to state sufficient conditions for the mapping $B(x)$ to be regular, similar to the conditions quoted in Section 1; e.g.,

1) $g_j(x, y) = D_j y + f_j(x)$, where D_j is a matrix; in the case, $B(x)$ is regular in any convex set X ;

2) $g_j(x, y)$ are continuous in $X \times Y$, and concave with respect to y , while Y is a convex compactum, and \tilde{y} exists, for which

$$\min_{1 \leq j \leq m} g_j(x^0, \tilde{y}) > 0;$$

then, $B(x)$ is regular at the point x^0 .

However, in the same way as in Section 1, examples can be quoted of non-concave $g_j(x, y)$, specifying regular mappings $B(x)$.

Lemma 1

If the many-valued mapping $B(x)$, specified by the continuous functions $g_j(x, y)$ and the compactum Y , is regular at the point x^0 , then it is Hausdorff - continuous at x^0 .

Proof. The upper semi-continuity of $B(x)$ is obvious. We shall prove the lower semi-continuity at the point x^0 . Assume that there exist $x_k \rightarrow x^0, y^k \in B(x^0), y^k \notin V_\delta(B(x_k))$ for $\delta > 0$. Then, by (7),

$$\min_{1 \leq j \leq m} g_j(x_k, y^k) \leq -K\delta < 0 \quad \forall k.$$

On the other hand, we can choose a subsequence $\{k_i\}$ such that $y^{k_i} \rightarrow y^0 \in B(x^0)$. Here,

$$\lim_{i \rightarrow \infty} \min_{1 \leq j \leq m} g_j(x_{k_i}, y^{k_i}) = \min_{1 \leq j \leq m} g_j(x^0, y^0) \geq 0,$$

and we arrive at a contradiction.

When $B(x)$ is regular in X , Lemma 1 guarantees the existence of an optimal strategy x^0 .

Theorem 3

Let

$$\frac{\partial}{\partial x} F(x, y), \quad \frac{\partial}{\partial y} F(x, y), \quad \frac{\partial}{\partial x} g_j(x, y), \quad \frac{\partial}{\partial y} g_j(x, y)$$

be continuous in the product of convex closed sets X, Y , where $X \subseteq E_n$, and Y is a bounded set of Euclidean space. Then there exist in problem (6) numbers $p_i \geq 0$ and numbers $\lambda_0, \lambda_{ij} \geq 0, 1 \leq$

$i \leq r \leq n+1, 1 \leq j \leq m$, not all vanishing, and also points y_i such that $\sum_{i=1}^r p_i = 1$,

$$-\sum_{i=1}^r p_i \left\{ \lambda_0 \frac{\partial}{\partial x} F(x^0, y_i) - \sum_{j=1}^m \lambda_{ij} \frac{\partial}{\partial x} g_j(x^0, y_i) \right\} \in K_X^*(x^0), \quad (8)$$

$$\left\{ \lambda_0 \frac{\partial}{\partial y} F(x^0, y_i) - \sum_{j=1}^m \lambda_{ij} \frac{\partial}{\partial y} g_j(x^0, y_i) \right\} \in K_Y^*(y_i),$$

$$\lambda_{ij} g_j(x^0, y_i) = 0, \quad (9)$$

$$y_i \in R(x^0) = \{y \in B(x^0) \mid F(x^0, y) = \min_{z \in B(x^0)} F(x^0, z)\}.$$

If the mapping $B(x)$ is regular at the point x^0 , then $\lambda_0 = 1$.

Proof. We shall use the scheme of arguments of Theorem 2. We introduce $F_1(x, y) = F(x, y) - \|x - x^0\|^2$, having the unique realization x^0 of the maximin

$$\sup_{x \in S \cap X} \min_{y \in B(x)} F_1(x, y).$$

By the theorem on penalty functions [4],

$$\sup_{x \in X \cap S} \min_{y \in B(x)} F_1(x, y) = \lim_{C \rightarrow \infty} \max_{x \in S \cap X} \min_{y \in Y} L_2(x, y, C),$$

where

$$L_2(x, y, C) = F_1(x, y) + C \sum_{j=1}^m [\min(0; g_j(x, y))]^2.$$

As $C_k \rightarrow \infty$, the realization of the maximin $x^0(C_k)$ of the function L_2 tends to x^0 . We write the necessary conditions for the point $x^0(C_k)$ [2]: there exist

$$\begin{aligned} p_i(C_k) &\geq 0 \text{ and } y_i(C_k) \in \tilde{R}(x^0(C_k)) = \{y \in Y \mid L_2(x^0(C_k), y, C_k) \\ &= \min_{z \in Y} L_2(x^0(C_k), z, C_k)\} \end{aligned}$$

such that

$$\sum_{i=1}^r p_i(C_k) = 1,$$

$$\begin{aligned} &-\sum_{i=1}^r p_i(C_k) \left\{ \frac{\partial}{\partial x} F_1(x^0(C_k), y_i(C_k)) \right. \\ &\left. - \sum_{j=1}^m s_{ij}(C_k) \frac{\partial}{\partial x} g_j(x^0(C_k), y_i(C_k)) \right\} \in K_X^*(x^0(C_k)). \end{aligned} \quad (10)$$

Since $y_i(C_k)$ realizes the minimum of $L_2(x^0(C_k), y, C_k)$, we have

$$\left\{ \frac{\partial}{\partial y} F_1(x^0(C_k), y_i(C_k)) - \sum_{j=1}^m s_{ij}(C_k) \frac{\partial}{\partial y} g_j(x^0(C_k), y_i(C_k)) \right\} \in K_Y^*(y_i(C_k)). \quad (11)$$

Here, $s_{ij}(C_k) = -2C_k \min(0; g_j(x^0(C_k), y_i(C_k)))$.

We normalize (1) and (11) to

$$m(C_k) = 1 + \sum_{i,j} s_{ij}(C_k) \geq 1,$$

by introducing

$$\lambda_0(C_k) = \frac{1}{m(C_k)}, \quad \lambda_{ij}(C_k) = \frac{s_{ij}(C_k)}{m(C_k)}.$$

Obviously, $\lambda_0(C_k), \lambda_{ij}(C_k)$ do not all vanish, since

$$\lambda_0(C_k) + \sum_{i,j} \lambda_{ij}(C_k) = 1.$$

We now pass to the limit in (10), (11) with respect to the subsequence $\{C_{k_l}\}$ in such a way that $y_i(C_{k_l}) \rightarrow y_i \in R(x^0)$, $\lambda_0(C_{k_l}) \rightarrow \lambda_0$, $\lambda_{ij}(C_{k_l}) \rightarrow \lambda_{ij}$.

We then obtain from (10), (11) the conditions (8), (9) and the conditions of supplementary non-rigidity.

If the mapping satisfies the regularity conditions at the point x^0 , we can easily show, in the same way as in Theorem 2, that the sums

$$\sum_{j=1}^m s_{ij}(C)$$

are uniformly bounded with respect to C . The conditions (8) and (9) here follow from (10) and (11), on passing to the limit as $C_{k_l} \rightarrow \infty$ without preliminary normalization.

It can easily be seen that the number of relations in the necessary conditions of Theorem 3 is equal to the number of unknown parameters, so that in principle the conditions contain sufficient information for finding x^0 .

In the case when Y is the entire space and the regular mapping $B(x)$ is bounded in the neighbourhood of x^0 , Theorem 3 may be stated in a different way. We introduce the Lagrange function

$$L(x, y, \lambda) = F(x, y) - \sum_{j=1}^m \lambda_j g_j(x, y), \quad \lambda_j \geq 0,$$

and the set

$$\Lambda(x^0, y) = \left\{ \lambda \geq 0 \mid \frac{\partial L(x^0, y, \lambda)}{\partial y} = 0, \quad \lambda g(x, y) = 0 \right\}.$$

Since the mapping $B(x)$ is regular at the point x^0 , and in view of Theorem 2, $\Lambda(x^0, y)$ will be non-empty for $y \in R(x^0)$.

Let Λ denote the set of functions $\lambda(\cdot)$, specified in $R(x^0)$ and such that $\lambda(y) \in \Lambda(x^0, y)$ for all $y \in R(x^0)$.

Corollary

The assertion of Theorem 3 is equivalent to the existence of $\lambda(\cdot) \in \Lambda$ such that

$$-M(\lambda(\cdot)) \cap K_X^*(x^0) = \emptyset. \quad (12)$$

Here,

$$M(\lambda(\cdot)) = \overline{\text{co}} \left\{ z \in E_n \mid z = \frac{\partial L(x^0, y, \lambda(y))}{\partial x}, y \in R(x^0) \right\},$$

where $\overline{\text{co}} A$ is the closure of the convex hull of the set A . From condition (12) we obtain the following necessary condition for the point x^0 :

$$\inf_{\lambda(\cdot) \in \Lambda} \sup_{\substack{\|g\|=1 \\ g \in K_X^*(x^0)}} \min_{y \in R(x^0)} \left(\frac{\partial L(x^0, y, \lambda(y))}{\partial x}, g \right) \leq 0 \quad (13)$$

(see Theorems 2 and 3 of [2]).

To check on condition (13), we need to solve a "non-classical" problem of seeking a min-max-min. In fact, the infimum in (13) is taken with respect to the set of functions $\lambda(\cdot)$, but nevertheless it is not possible to write (13) in the form

$$\sup_{\substack{\|g\|=1 \\ g \in K_X^*(x^0)}} \min_{y \in R(x^0)} \min_{\lambda \in \Lambda(x^0, y)} \left(\frac{\partial L(x^0, y, \lambda)}{\partial x}, g \right) \leq 0,$$

i.e., to go over to a minimax problem in Euclidean space. This suggests the idea that the necessary conditions are more convenient to use in the form (12), rather than in the form (13). The same conclusion holds for the problem of seeking a maximin with splitting variables [2, 3].

Notice that problems of the type (13) have recently appeared with increasing frequency in connection with the analysis of hierarchical control system.

It was shown in [2] (Theorem 4.2) that, under more rigid assumptions, ensuring the directional differentiability of the minimum function

$$f(x) = \min_{y \in B(x)} F(x, y)$$

conditions (12) hold for any function $\lambda(\cdot) \in \Lambda$. In this case, the analogue of condition (13) is

$$\sup_{\lambda(\cdot) \in \Lambda} \sup_{\substack{\|g\|=1 \\ g \in K_X^*(x^0)}} \min_{y \in R(x^0)} \left(\frac{\partial L(x^0, y, \lambda(y))}{\partial x}, g \right) \leq 0,$$

which, in view of the fact that \sup and \min permute, is equivalent to

$$\sup_{\substack{\lambda(\cdot) \in \Lambda \\ \|g\|=1 \\ g \in K_X(x^0)}} \min_{y \in R(x^0)} \max_{\lambda \in \Lambda(x^0, y)} \left(\frac{\partial L(x^0, y, \lambda)}{\partial x}, g \right) \leq 0,$$

i.e., an ordinary maximin problem.

3. Sequential maximin with constraints

Let x_1^0 realize

$$\begin{aligned} & \max_{x_1 \in A_1} \min_{y_1 \in B_1(x_1)} \dots \\ & \dots \max_{x_n \in A_n(x_1, y_1, \dots, y_{n-1})} \min_{y_n \in B_n(x_1, \dots, x_n)} \max_{x_{n+1} \in A_{n+1}(x_1, \dots, y_n)} F(x_1, y_1, \dots, y_n, x_{n+1}) = M, \end{aligned} \quad (14)$$

or more briefly,

$$\begin{aligned} M &= \left[\max_{x_i \in A_i} \min_{y_i \in B_i} \right]_{i=1}^n \max_{x_{n+1} \in A_{n+1}} F(x_1, y_1, \dots, y_n, x_{n+1}) \\ &= \left[\min_{y_j \in B_j} \max_{x_{j+1} \in A_{j+1}} \right]_{j=1}^n F(x_1^0, y_1, \dots, x_n, y_n, x_{n+1}). \end{aligned} \quad (15)$$

Here, the mappings A_i and B_j are specified in the form

$$\begin{aligned} A_i &= A_i(x_1, \dots, y_{i-1}) = \{x_i \in X_i \mid h_{il}(x_i, y_1, \dots, y_{i-1}, x_i) \geq 0\}, \\ B_j &= B_j(x_1, \dots, x_j) = \{y_j \in Y_j \mid g_{jm}(x_1, y_1, \dots, x_j, y_j) \geq 0\}, \end{aligned}$$

$1 \leq i \leq n+1$, $1 \leq j \leq n$, where l and m run over finite sets of indices.

We shall make the following assumptions:

1) the sets x_p , Y_j are convex and closed, and belong to r_i - and s_j -dimensional Euclidean spaces respectively;

2) the mappings A_i and B_j are bounded and Hausdorff-continuous;

3) the functions h_{il} , g_{jm} have continuous partial derivatives.

We introduce the Lagrange function connected with the problem (14):

$$\begin{aligned} & L(x_1, y_1, \dots, x_n, y_n, x_{n+1}; \lambda_0, \lambda_1, \dots, \lambda_n; \mu_1, \dots, \mu_{n+1}) \\ &= \lambda_0 F(x_1, \dots, y_n, x_{n+1}) - \sum_{j=1}^n (\lambda_j, g_j(x_1, y_1, \dots, y_j)) \\ &+ \sum_{i=1}^{n+1} (\mu_i, h_i(x_1, y_1, \dots, x_i)), \end{aligned} \quad (16)$$

$$\begin{aligned}
&= \max_{x_{n+1} \in A_{n+1}} F(x_1^0, y_1^{(i_1)}, x_2^{(i_1, i_2)}, \dots, x_{n+1}) \\
&= F(x_1^0, y_1^{(i_1)}, x_2^{(i_1, i_2)}, \dots, x_{n+1}^{(i_1, i_2, \dots, i_n)}).
\end{aligned}
\tag{cont'd}$$

If the mappings A_i, B_j are regular in the product of the relevant X_i and Y_j , then the coefficient $\lambda_0 = 1$ in the Lagrange form (16).

Proof. By the convergence theorem for the penalty method in problem (14) [4]

$$M = \lim_{C \rightarrow \infty} [\max_{x_i \in X_i} \min_{y_i \in Y_i}]_{i=1} \max_{x_{n+1} \in X_{n+1}} L(x_1, y_1, \dots, x_n, y_n, x_{n+1}, C),$$

where

$$\begin{aligned}
L(x_1, y_1, \dots, x_{n+1}, C) &= F(x_1, y_1, \dots, y_n, x_{n+1}) \\
&+ C \sum_{j=1}^n \sum_{m_i} [\min(0; g_{jm_i}(x_1, \dots, y_i))]^2 \\
&- C \sum_{i=1}^{n+1} \sum_l [\min(0; h_{il}(x_1, \dots, x_i))]^2.
\end{aligned}$$

On writing here the necessary conditions for optimality for the multiple maximin [4], then passing to the limit as $C \rightarrow \infty$, we obtain the theorem.

It would seem that the most important regular case of problem (14) corresponds to linear connections, when

$$\begin{aligned}
A_i &= \left\{ x_i \in X_i \mid \sum_{h=1}^i E_{hi} x_h + \sum_{t=1}^{i-1} D_{ti} y_t \leq a_i \right\}, \\
B_j &= \left\{ y_j \in Y_j \mid \sum_{h=1}^j H_{hj} x_h + \sum_{t=1}^j G_{tj} y_t \leq b_j \right\},
\end{aligned}$$

where $E_{hi}, D_{ti}, H_{hj}, G_{tj}$ are matrices, and a_i, b_j are vectors of the respective spaces.

Translated by D. E. Brown

REFERENCES

1. DEM'YANOV, V. F., On the minimax problem with connected constraints, *Zh. vychisl. Mat. mat. Fiz.*, **12**, No. 3, 799-804, 1972.
2. DEM'YANOV, V. F., *The minimax: directional differentiability* (Minimaks: differentsiruemost' po napravleniyam), Izd-vo LGU, Leningrad, 1974.
3. GERMEIER, YU. B., The constrained max-min problem, *Zh. vychisl. Mat. mat. Fiz.*, **10**, No. 1, 39-54, 1970.
4. FEDOROV, V. V., *Methods of seeking a maximin* (Metody poiska maksimuma), Izd-vo MGU, Moscow, 1975.
5. DUBOVITSKII, A. YA., and MILYUTIN, A. A., Constrained extremum problems, *Zh. vychisl. Mat. mat. Fiz.*, **5**, No. 3, 395-453, 1965.

6. IOFFE, A. D., and TIKHOMIROV, V. M., *Theory of extremal problems* (Teoriya ekstremal'nykh zadach), Nauka, Moscow, 1974.
7. GOL'SHEIN, E. G., *Theory of duality in mathematical programming and its applications* (Teoriya dvoistvennosti v matematicheskoy programirovani i ee prilozheniya), Nauka, Moscow, 1971.
8. EREMIN, I. I., On the method of "penalties" in convex programming, *Kibernetika*, No. 4, 63-67, 1967.

A STOCHASTIC QUASI-GRADIENT METHOD FOR SEEKING A MAXIMIN*

N. M. NOVIKOVA

Moscow

(Received 24 April 1975; revised 10 November 1975)

THE ALGORITHM described below for seeking a maximin is a combination of the penalty method and the stochastic quasi-gradient method. A theorem on convergence to the set of solutions with probability unity is proved. The ALGOL program is given, along with test results.

1. Consider the problem of finding

$$u^0 = \max_{x \in X} \min_{y \in Y} f(x, y) \quad (1)$$

and the best guaranteeing strategy $x^0 \in X$:

$$\min_{y \in Y} f(x^0, y) = u^0. \quad (2)$$

We assume that the function $f(x, y)$ is continuous with respect to x and y in $X \subset E^p$ and $Y \subset E^m$ which are closed bounded sets of Euclidean space. In accordance with [1], the problem is equivalent to finding $[u - \alpha_n \Phi_q(x, u)]$ with respect to x, u in the set $X \times [M_1, M_2]$ as $\alpha_n \uparrow +\infty$, where

$$\Phi_q(x, u) = \int_Y |\min[0; f(x, y) - u]|^q \sigma(dy), \quad 1 \leq q \leq 2,$$

$$M_1 = \min_{X \times Y} f(x, y), \quad M_2 = \max_{X \times Y} f(x, y).$$

Putting $F_n^q(x, u) = u - \alpha_n \Phi_q(x, u)$, we have

$$u^0 = \lim_{n \rightarrow \infty} \max_{X \times [M_1, M_2]} F_n^q(x, u) = \lim_{n \rightarrow \infty} F_n^q(x_n^0, u_n^0); \quad (1')$$

the vector (x_n^0, u_n^0) realizes

$$\max_{X \times [M_1, M_2]} F_n^q(x, u),$$

any limit point of the sequence $\{(x_n^0, u_n^0)\}$ being a solution of problem (1), (2). In other words, $\{(x_n^0, u_n^0)\}$ is convergent to $X^0 \times u^0$, where $X^0 = \{x^0\}$ is the set of solutions of Eq. (2).

In [2], under the extra assumption that $f(x, y)$ satisfies a Lipschitz condition with respect to y for all $x \in X$, the following convergence rate estimate is obtained for the method of penalties (1'):

$$0 \leq u_n^0 - u^0 \leq D(1/\alpha_n)^{1/(m+q-1)}, \quad D = \text{const},$$

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 91-99, 1977.

for sufficiently large u . If we also recall (see [2]) that $u^0 \leq F_n^q(x_n^0, u_n^0) \leq u_n^0$, we get $0 \leq F_n^q(x_n^0, u_n^0) - u^0 \leq D(1/\alpha_n)^{1/(m+q-1)}$ and since $\Phi_q(x^0, u^0) = 0$, we have

$$0 \leq F_n^q(x_n^0, u_n^0) - F_n^q(x^0, u^0) \leq D(1/\alpha_n)^{1/(m+q-1)}. \quad (3)$$

Gradient methods* are usually employed to find x_n^0, u_n^0

To avoid the difficulties involved in evaluating the integral in $\text{grad } \Phi_q(x, u)$, we use the stochastic quasi-gradient method (see [3]). Let $\int \sigma(dy) = 1$ (appropriate normalization); for instance, let Y be the m -dimensional unit cube, and σ the ordinary Lebesgue measure. The integral in question may then be interpreted as the mathematical expectation of the integrand of the random variable y , subject to a uniform distribution law in Y . On randomly choosing $y_n \in Y$ (with equal probability), we can construct the random sequence $\{(x_n, u_n)\}$, convergent almost surely (with probability unity) to $X^0 \times u^0$, i.e., the set of solutions of the problem (1), (2).

We introduce $\chi_n^q(y|x, u) = u - \alpha_n |\min[0; f(x, y) - u]|^q$ is a function of the random variable y , dependent on x, u , whose mathematical expectation

$$M_y \{\chi_n^q(y|x, u)\} = F_n^q(x, u). \quad (4)$$

In future we shall seek the indices of the mathematical expectation. We assume that X is convex, and $f(x, y)$ is concave with respect to x , i.e., $\chi_n^q(y|x, u)$ is concave with respect to (x, u) . We denote the vector (x, u) by z ; the set $Z = X \times [M_1, M_2]$ is convex. We define $\xi_n^q(y|x, u)$ as the generalized gradient with respect to x, u for fixed y , of the function $\chi_n^q(y|x, u)$:

$$\langle \xi_n^q(y|z^1), z^1 - z^2 \rangle \leq \chi_n^q(y|z^1) - \chi_n^q(y|z^2) \quad \forall z^1, z^2 \in Z \quad (5)$$

(here, \langle, \rangle is the scalar product). For instance, in the case of a discontinuous penalty, on taking $q = 1$, we can choose for $p = 1$ the components of the vector $\xi_n^1(y|x, u) = (\xi_n(y|x, u); \eta_n(y|x, u))$ as follows:

$$\begin{aligned} \xi_n(y|x, u) &= -\alpha_n \frac{\partial f^-(x, y)}{\partial x} \text{sign}[\max(0; f(x, y) - u)], \\ \eta_n(y|x, u) &= 1 + \alpha_n \text{sign}[\max(0; f(x, y) - u)], \end{aligned}$$

since the concave function is differentiable with respect to any direction, the right- or left-hand partial derivative (say the left-hand, $\partial f^+(x, y)/\partial x$ or $\partial f^-(x, y)/\partial x$) will exist. Since the sets Y, Z are bounded, the following estimate for the Euclidean norm of the generalized gradient is obvious: $\|\xi_n^q(y|z)\|^2 \leq C\alpha_n^2$, $C = \text{const}$ (since the finite generalized gradient of $f(x, y)$ with respect to x exists).

Under our assumptions, condition (4) allows us to apply the result of [3] to problem (1'). Using this result, we can construct random sequences $\{z_n^k\}_{k=1}^\infty$, convergent almost surely to the sets $\{z_n^0\}$ respectively. Then,

$$u^0 = \lim_{n \rightarrow \infty} \lim_{k \rightarrow \infty} \chi_n^q(y_n^k|z_n^k)$$

*The method of gradient projections (generalized gradient method).

almost surely; $\{y_n^k\}_{k=1}^\infty$ are sequences of independent, uniformly distributed random variables for any $n = 1, 2, \dots$. However, the use of the repeated limit is inconvenient when it comes to practical computation. It can be avoided by the method described below, which takes account of the specific nature of the problem.

Assume that the numerical sequences $\{a_n\}$, $\{\alpha_n\}$ are such that the following conditions are satisfied:

$$a_n \geq 0, \quad a_n \downarrow 0, \quad \sum_{n=1}^{\infty} a_n = +\infty, \quad \alpha_n \uparrow +\infty, \quad (6)$$

$$\sum_{n=1}^{\infty} a_n^2 \alpha_n^2 < \infty, \quad \sum_{n=1}^{\infty} a_n \left(\frac{1}{\alpha_n} \right)^{1/(m+q-1)} < \infty. \quad (7)$$

We specify arbitrary $z_1 \in Z$ and define $\{z_n\}_{n=1}^\infty$ as follows

$$z_{n+1} = \pi \{ z_n + a_n \xi_n^q(y_n | z_n) \}, \quad y_n \in Y; \quad (8)$$

$\pi \{ \}$ is the projection onto Z , and $\{y_n\}_{n=1}^\infty$ is a sequence of independent, uniformly distributed random variables.

Theorem

Under our assumptions, the probability of $\{z_n\}$ being convergent to $Z^0 = X^0 \times u^0$, is unity, i.e., the distance $\rho(z_n, Z^0)$ between z_n and Z^0 tends to zero almost surely.

Proof. Using (8) and (5), we have

$$\begin{aligned} \|z_{n+1} - z^0\|^2 &\leq \|z_n + a_n \xi_n^q(y_n | z_n) - z^0\|^2 = \|z_n - z^0\|^2 \\ &+ 2a_n \langle \xi_n^q(y_n | z_n), z_n - z^0 \rangle + a_n^2 \|\xi_n^q(y_n | z_n)\|^2 \\ &\leq \|z_n - z^0\|^2 + 2a_n [\chi_n^q(y_n | z_n) - \chi_n^q(y_n | z^0)] + Ca_n^2 \alpha_n^2. \end{aligned}$$

Consequently, since $z^0 \in Z^0$ is arbitrary, we obtain

$$\min_{z^0 \in Z^0} \|z_{n+1} - z^0\|^2 \leq \min_{z^0 \in Z^0} \|z_n - z^0\|^2 + 2a_n [\chi_n^q(y_n | z_n) - u^0] + Ca_n^2 \alpha_n^2,$$

since $\chi_n^q(y_n | z_n^0) = u^0 \quad \forall y_n \in Y$;

$$\begin{aligned} M \{ \min_{z^0 \in Z^0} \|z_{n+1} - z^0\|^2 | z_1, \dots, z_n \} &\leq \min_{z^0 \in Z^0} \|z_n - z^0\|^2 + Ca_n^2 \alpha_n^2 \\ &+ 2a_n [F_n^q(z_n) - u^0] \leq \min_{z^0 \in Z^0} \|z_n - z^0\|^2 + 2a_n [F_n^q(z_n^0) - u^0] + Ca_n^2 \alpha_n^2, \end{aligned}$$

since

$$F_n^q(z_n) \leq F_n^q(z_n^0) = \max_z F_n^q(z);$$

the previous inequalities (4) and the properties of the conditional mathematical expectation have also been used.

Hence from (3) we have

$$\begin{aligned} \mathbf{M}\{\rho^2(z_{n+1}, Z^0) | z_1, \dots, z_n\} &\leq \rho^2(z_n, Z^0) + Ca_n^2 \alpha_n^2 \\ &\quad + 2Da_n(1/\alpha_n)^{1/(m+q-1)}, \end{aligned}$$

i.e., the sequence $\{\rho^2(z_n, Z^0)\}_{n=1}^\infty$ is convergent with probability unity. For, since (see [3]) the sequence

$$\{w_n\}_{n=1}^\infty, \quad w_n = -\rho^2(z_n, Z^0) - \sum_{k=n+1}^\infty [2Da_k(1/\alpha_k)^{1/(m+q-1)} + Ca_k^2 \alpha_k^2]$$

forms a semi-martingale (as is clear from (7)), the required convergence will follow from the properties of semi-martingales (see [4], Chapter 7, Section 4, Theorem 4.1 (1)).

After this, on taking the unconditional mathematical expectation of both sides of the inequalities and performing the summation, we obviously get

$$\begin{aligned} \forall z^0 \in Z^0 \quad 0 \leq \mathbf{M}\{\|z_{n+1} - z^0\|^2\} &\leq \|z_1 - z^0\|^2 + C \sum_{k=1}^n a_k^2 \alpha_k^2 \\ &\quad + 2 \sum_{k=1}^n a_k \mathbf{M}\{F_k^q(z_k) - F_k^q(z^0)\}, \quad n=1, 2, \dots \end{aligned}$$

This follows from the properties of the mathematical expectation (the unconditional expectation of the conditional expectation is equal to the unconditional expectation). Hence

$$\sum_{k=1}^\infty a_k \mathbf{M}\{F_k^q(z_k) - F_k^q(z^0)\} > -\infty,$$

since

$$\sum_{k=1}^\infty a_k^2 \alpha_k^2 < \infty.$$

By definition, $F_n^q(z^0) = u^0$ for $z^0 \in Z^0$; hence

$$\sum_{n=1}^\infty a_n \mathbf{M}\{F_n^q(z_n) - u^0\} > -\infty.$$

Consequently,

$$\sum_{n=1}^\infty a_n \mathbf{M}|F_n^q(z_n) - u^0| < \infty,$$

since, in addition to the previous inequality,

$$\sum_{n=1}^\infty a_n \mathbf{M}\{\max[0; F_n^q(z_n) - u^0]\} < \infty,$$

inasmuch as

$$\mathbf{M}\{\max[0; F_n^q(z_n) - u^0]\} \leq D(1/\alpha_n)^{1/(m+q-1)},$$

which in turn follows from (3) (the measure of Y is equal to 1). Then, from conditions (6) on the sequence $\{a_n\}$ we have

$$\lim_{n \rightarrow \infty} M\{|F_n^q(z_n) - u^0|\} = 0.$$

Hence a subsequence $\{F_{n_k}^q(z_{n_k})\}_{k=1}^{\infty}$ exists, convergent to u^0 almost surely.

We are now able to show that $\{z_n\}$ is convergent almost surely to Z^0 , i.e., $\rho(z_n, Z^0) \rightarrow 0$ almost surely as $n \rightarrow \infty$. For, denote by A the set

$$\{\omega = \{y_n\}_{n=1}^{\infty} | F_{n_k}^q(z_{n_k}(\omega)) \xrightarrow[k \rightarrow \infty]{} u^0, \rho^2(z_n(\omega), Z^0) \text{ converges}\}.$$

By what has been proved above, the probability measure of this set $P\{A\} = 1$.

Since Z is bounded, for all $\omega \in A$ the sequence $\{z_{n_k}(\omega)\}$ has a convergent subsequence $\{z_{n_{kp}(\omega)}(\omega)\}$, henceforth denoted by $\{z_{np}(\omega)\}$:

$$z_{n_p}(\omega) \xrightarrow[p \rightarrow \infty]{} \hat{z}(\omega), \quad F_{n_p}^q(z_{n_p}(\omega)) \xrightarrow[p \rightarrow \infty]{} u^0(\omega) \equiv u^0.$$

In other words,

$$u_{n_p}(\omega) - \alpha_{n_p} \Phi_q(z_{n_p}(\omega)) \xrightarrow[p \rightarrow \infty]{} u^0.$$

Hence $\alpha_{n_p} \Phi_q(z_{n_p}(\omega))$ is bounded (since u_{n_p} and Z are bounded), and

$$\Phi_q(z_{n_p}(\omega)) \xrightarrow[p \rightarrow \infty]{} 0 \quad \forall \omega \in A, \quad \text{since } \alpha_{n_p} \rightarrow +\infty.$$

It then follows from the continuity of $\Phi_q(z)$ that $\Phi_q(\hat{z}(\omega)) = 0$ for all $\omega \in A$. Hence,

$$\hat{u}(\omega) \leq \min f(\hat{x}(\omega), y).$$

In addition, $\hat{u}(\omega) \geq u^0$ (since $u_{n_p}(\omega) \rightarrow \hat{u}(\omega)$, $u_{n_p}(\omega) - \alpha_{n_p} \Phi_q(z_{n_p}(\omega)) \rightarrow u^0$, $\alpha_{n_p} \Phi_q(z_{n_p}(\omega)) \geq 0$). Hence

$$\min_Y f(\hat{x}(\omega), y) \geq \hat{u}(\omega) \geq u^0 = \max_X \min_Y f(x, y)$$

and since $\hat{z}(\omega) \in Z$ (where Z is closed), we have $\hat{x}(\omega) = x^0(\omega)$, $\hat{u}(\omega) = u^0$, i.e., $\hat{z}(\omega) \in Z^0$ for all $\omega \in A$. Hence $\rho(z_{n_p}(\omega), Z^0) \rightarrow 0$ as $p \rightarrow \infty$ and $\rho^2(z_n(\omega), Z^0)$ are convergent; we now finally have

$$\rho(z_n(\omega), Z^0) \xrightarrow[n \rightarrow \infty]{} 0 \quad \forall \omega \in A.$$

Since $P\{A\} = 1$, the proof of the theorem is complete.

The theorem can obviously be restated as follows: $\{u_n\}$ is convergent to u^0 almost surely, and any limit point of the sequence $\{x_n\}$ is some $x^0 \in X^0$ with probability unity.

2. Let us now make some general remarks concerning the proposed method. Its merits are simplicity of realization, and reduced demand on the smoothness and stability of the computational errors. For, in order to satisfy the conditions of the theorem, it is sufficient that

$$M\{\zeta_n^q(y_n | z_n) | y_n, z_n\} = \text{grad}_x \chi_n^q(y_n | z_n) + r_n, \quad \sum_{n=1}^{\infty} a_n |r_n| < \infty \quad (9)$$

(here, grad_z denotes the generalized gradient with respect to z). This enables us to use approximate methods to evaluate the generalized gradient, i.e., to find $\text{grad}_x f(x_n, y_n)$ we can use $\hat{f}_x(x_n, y_n)$:

$$M\{\hat{f}_x(x_n, y_n) | x_n, y_n\} = \text{grad}_x f(x_n, y_n) + s_n, \quad \sum_{n=1}^{\infty} \alpha_n \alpha_n |s_n| < \infty.$$

The conditions obtained are similar to the requirements of [3].

With regard to the rate of convergence, the computation of test examples showed that a first approximation to u^0 , and quite a short distance from x^0 , can be reached fairly rapidly; but further improvement of u^0 takes a long time, so that certain modifications of the algorithm can be recommended for accelerating the convergence in practice (see Appendix). Of course the computer time increases in proportion to the dimensionality of the problem.

Of course the method can be used, not only when seeking a maximin, but also when maximizing a concave function under general constraints:

$$f(z^0) = \max_{z \in G} f(z), \quad G = \{z \in Z | g(z, y) \geq 0 \quad \forall y \in Y\}, \quad (10)$$

if it is assumed that the function $g(x, y)$ is concave with respect to $z \in Z$, where Z is a convex compactum.

In this case the function

$$F_n^q(z) = f(z) - \alpha_n \int_Y |\min[0; g(z, y)]|^q dy$$

is concave with respect to z . Further, if z_n^0 realizes $\max_z F_n^q(z)$, then it follows from the method of penalties (see [1]) that

$$\lim_{n \rightarrow \infty} F_n^q(z_n^0) = f(z^0)$$

and any limit point of the sequence $\{z_n^0\}_{n=1}^{\infty}$ proves to be a solution of problem (10). On using the estimates of [2] (which are obtained for finite Y under rigid conditions on $g_i(z)$), we can choose a_n and α_n in such a way that the algorithm (8) converges with probability unity to the set $\{z^0\}$ in the conditions demanded in [2]. Use may also be made of other estimates of the convergence of $F_n^q(z_n^0)$ to $f(z^0)$.

The algorithm (8) can obviously be used when Y is a finite set. In particular, if Y consists of a single point, i.e., we are concerned with an ordinary constrained extremum problem, where the error is of the type (9), we obtain convergence with probability unity (when there are no errors, the convergence is guaranteed). This remark has something in common with the result obtained in [5] for penalties of a different kind, under more rigid smoothness conditions.

Appendix ALGOL program and tests

The program written below, as a procedure approx in ALGOL 60, realizes the algorithm (8) (with provision for possible modifications in the case $0 < l \leq 1$) in the case when X is the p -dimensional cube, Y is the m -dimensional cube, $x^i \in [0, 1]$, $i=1, 2, \dots, p$, $y^j \in [0, 1]$, $j=1, 2, \dots, m$,

and $f(x, y)$ varies between 0 and 2. The parameters x, u of the procedure correspond to the running value x_n, u_n , and prior to access to the procedure they have to be given their initial values. The control constants $i0, i1$ are the initial and final values of n ; $j0$ is the step in n . After $j1$ steps the values of x, u are printed by means of the procedure output (x, u) , which can be replaced by another standard output procedure (according to the type of computer). The constants r, s specify the order of decrease of a_n, a_n, α_n (for instance, $r = 1, s = 0.6$). The procedure-function $\text{func}(p, m, x, y)$ has the value $f(x, y)$. The procedure $\text{grad}(p, m, x, y, g)$ has to evaluate the generalized gradient of $f(x, y)$ with respect to x and assign it to g . The parameter $l, 0 < l \leq 1$, serves for different modifications of the method; the value $l > 1/2$ is used after obtaining the first approximation for sufficiently large $i0, 1 \leq q \leq 2$. The procedure-function rand in the body of the procedure is a source of random numbers, uniformly distributed in the interval $[0, 1]$; for concrete translators it can be replaced by a standard procedure (e.g., $p1147(a1, \text{rand})$ for the TA-1M); in this case the initial assignment to the variables $a0, a1$ has to be replaced by initial access to the standard procedure ($p1147(a1)$ for TA-1M).

The approx procedure is best used several times, using the approximations obtained as the initial approximations, while increasing $i0, i1, j1$, until after a sufficient number of steps the upper bound of fluctuation for u is stabilized, and the value of x is established with the necessary accuracy*. If the fluctuation of u are sparse, $j0$ can be increased.

In addition to the modifications provided for in the procedure (with $l \neq 0$), it is possible, in certain problems which demand that u^0 be determined with increased accuracy, to introduce a variable upper bound for u for subsequent refinements. At the initial instant it is assigned the value M_2 , while later in the cycle it is re-assigned the value u , which does not satisfy the constraints (for $h < 0$), and at each step u is compared, not with M_2 (in the procedure, $M_2 = 2$), but with the variable value introduced. We have to see to it here that the amount of variation of x is within the given accuracy range, since otherwise the value M_2 must again be assigned to the bound for u . However, in problems where 5% accuracy is sufficient, there is usually no point in making our procedure more complicated. Some examples of rationalization of the procedure will be given, along with the tests.

```

procedure approx (p, m, x, u, i0, i1, j0, j1, r, s, l, q, func, grad);
value p, m; untether p, m, i0, i1, j0, j1; array x;
real u, r, s, l, q; real procedure func; procedure grad;
begin integer i, j, k; real a0, a1, f, h, h1; array g [1 : p], y [1 : m];
  real procedure rand;
  begin real a; a := a0 + a1; a0 := a1;
    if a ≥ 4 then a := a - 4; a1 := a; rand := a/4
  end rand;
  a0 := 3.14159265; a1 := 0.542101887; h := 0;
  for i := i0 step j1 until i1 - j1 do
    begin for j := 1 step j0 until j1 do
      begin for k := 1 step 1 until m do y [k] := rand;
        f := func (p, m, x, y);
        h1 := if f ≥ u then 0 else -(u - f) ↑ (q - 1);
        if h1 ≤ h then h := h1 else h := l × h + (1 - l) × h1;
        u := u + 1 / (i + j) ↑ r + h / (i + j) ↑ s;
        if u < 0 then u := 0; if u > 2 then u := 2;

```

*In the case when X^0 consists of more than one point, the variation of x is checked for cycling.

```

grad (p, m, x, y, g);
for k:=1 step 1 until p do
begin x[k]:=x[k]-g[k]×hI/(i+j)↑s; if x[k]<0 then x[k]:=0;
  if x[k]>1 then x[k]:=1
end
end; output (x, u)
end
end approx

```

(cont'd)

The test examples were computed on the BESM-4 computer for the parameters $a_n=1/n$, $\alpha_n=n^{2/5}$ (for $a_n=a/n$ it is advisable to choose a of the order of M_2). The computational results are given in Table 1.

In Example I: $f(x, y)=1+(x-1/2)(y-1/2)$, the exact solution is $x^0=1/2$, $u^0=1$. In Example II, $f(x, y)=\exp(-(x^1-y^2)^2)+\exp(-(x^2-y^2)^2)$; here $x=(x^1, x^2)$ is a two-dimensional vector, and $x^0=(1/2, 1/2)$, $u^0=1.55$. In Example III, $x=(x^1, x^2)$, $y=(y^1, y^2)$, $f(x, y)=\exp(-(x^1-y^1)^2-(x^2-y^2)^2)$, $x^0=(1/2, 1/2)$, $u^0=0.606$.

The t column gives the computing time in minutes.

TABLE 1

Examples	x_1	u_1	$i0$	$i1$	$j0$	$j1$; i step	l	x_{i1}	u_{i1}	t	Modifi- cations
I {	0	0	0	10^4	1	50	$1/2$	0.500	1.007	5	
	$1/2$	1	10^4	10^5	10	500	$1/2$	0.500	1.002	5	
II {	(0, 0)	0	0	10^4	1	50	$1/2$	0.504	1.66	10	
	(0, 0)	0	0	10^5	10	500	$\frac{i}{i+1}$	0.504 0.497	1.58	13	(11)
	$(1/2, 1/2)$	1	0	10^4	1	50	1	0.494 0.494	1.57	10	
	$(1/2, 1/2)$	1.56	0	10^5	5	200	$\frac{i}{i+1}$	0.4999 0.4999	1.56	20	(11)
	$(1/2, 1/2)$	1.56	10^5	10^6	5	$i/200$; 5000	1	0.5000 0.5000	1.55	45	(12)
	(0, 0)	0	$0, 10^4$	10^4 , 10^5	5	200; 50, 500	$\frac{i}{i+1}$	0.50 0.49	0.62	25	(13), (11)
III {	$(1/2, 1/2)$	0.61	10^5	10^6	5	$i/200$; 2000	1	0.49999 0.50000	0.608	45	(12)

We also indicate the practical modifications of the approx procedure that were employed. For instance, the parameter l can be described as a variable, and not as a constant, real procedure l , and specified in the body of the approx procedure as follows:

real procedure l ; begin $l:=i/(i+1)$ end; (11)

The description of $j1$ for the variable end of the j cycle can be similarly modified:

real procedure $j1$; begin $j1:=i/200$ end; (12)

here, the i step must naturally remain constant and either be specified directly by a number, or denoted by a new identifier. It is also convenient to use a composite i cycle of the type

for $i:=0$ step 50 until 10^4 , 10^4 step 500 until 10^5 (13)

The author thanks Yu. B. Germeier for his interest, and V. V. Fedorov for guidance and practical advice.

Translated by D. E.

REFERENCES

1. GERMEIER, YU. B., Approximate reduction of the maximin problem to a maximum problem by means of penalty functions, *Zh. vychisl. Mat. mat. Fiz.*, 9, No. 3, 730-731, 1969.
2. FEDOROV, V. V., On the method of penalty functions in the problem of finding a maximin, *Zh. vychisl. Mat. mat. Fiz.*, 12, No. 2, 321-333, 1972.
3. ERMOL'EV, YU. M., On the method of generalized stochastic gradients in stochastic quasi-Fejer sequences, *Kibernetika*, No. 2, 73-83, 1969.
4. DOOB, J., *Stochastic processes*, Wiley, 1953.
5. FABIAN, V., Stochastic approximation of constrained minima, *Trans. 4-th Prague Conf. Inf. Theory, Statistic. Decis. Functions, Random Proc.*, 1965, Publ. House Czechosl. Acad. Sci., Prague, 277-290, 1967.

AN ITERATIVE METHOD WITH CHEBYSHEV PARAMETERS FOR FINDING THE MAXIMUM EIGENVALUE AND CORRESPONDING EIGENFUNCTION*

V. I. LEBEDEV

Moscow

(Received 6 January 1975; revised 12 July 1976)

TO ACCELERATE the convergence of the iterations when finding the maximum eigenvalue and corresponding eigenfunction, a method is proposed which employs infinite sequences of Chebyshev parameters and guarantees stability of the computations. A generalization of Bernoulli's method is constructed.

A typical example of the class of problems, in which an iterative method which accelerates the convergence of the iterations is used, is the difference analogue of the boundary value problem in a domain D with boundary Γ for the many-group diffusion equations [1]:

$$-\operatorname{div} D_i \operatorname{grad} \varphi_i + \Sigma_i \varphi_i = \sum_{j=1}^n \Sigma_i^{ij} \varphi_j + \frac{1}{\lambda} \chi_i Q \varphi_i, \quad (1)$$

$$d_i \frac{\partial \varphi_i}{\partial n} + \varphi_i|_{\Gamma} = 0, \quad (2)$$

**Zh. vychisl. Mat. mat. Fiz.*, 17, 1, 100-108, 1977.

where

$$Q\varphi = \sum_{j=1}^n v_{\Sigma_{ij}} \varphi_j. \quad (3)$$

It is required to find the least value of λ and the corresponding eigenfunction $(\varphi_1, \dots, \varphi_n)$. We write the problem in the operator form

$$L\varphi = \frac{1}{\lambda} \chi Q\varphi. \quad (4)$$

Assuming that the operator L^{-1} exists, putting $x = Q\varphi$, and $A = QL^{-1}\chi$ and applying to both sides of (4) the operator QL^{-1} , we get

$$Ax = \lambda x. \quad (5)$$

We propose to dwell on the properties of the operators in (4) which were utilized when constructing our iterative method.

Property 1. Finding the element u representing the solution of the equation $Lu = y$ involves a fairly laborious iterative method, which includes both interior (for each i) and exterior (with respect to i) cycles of iterations.

Property 2. The simultaneous storage in the computer memory of the values of the ϕ_i for several iterations demands very large capacity, while storage of $Q\phi$ is not so difficult.

Property 3. The eigenvalues of problem (4) will be assumed to be positive; from "physical considerations", a lower bound can be set for the maximum eigenvalue.

Property 4. The two largest eigenvalues may be close to one another, i.e., the method of simple iteration may converge slowly.

Property 5. As a rule, the maximum eigenvalue is found to a given accuracy much faster than the eigenfunction; to find the latter, special methods for speeding up the convergence have to be used, based on information obtained when finding the eigenvalue.

1. Iterative method

Let A be a linear bounded operator, specified in Banach space B and having a complete linearly independent system of normalized eigenelements $\varphi_1, \dots, \varphi_n, \dots$, corresponding to the eigenvalues $\lambda_1 > \lambda_2 \geq \dots \geq \lambda_n \geq \dots \geq 0$, where $\lambda_n \rightarrow 0$ as $n \rightarrow \infty$, and assume that a quantity $0 \leq a < \lambda_1$ can be specified *a priori*. Let $l(x)$ be a linear functional of the adjoint space B^* , $l_n = l(\varphi_n)$, where $l_1, l_2 \neq 0$. It is required to find the maximum eigenvalue λ_1 and the eigenfunction ϕ_1 of problem (5).

We shall use an iterative method with variable displacement, the size of which will be determined by an infinite sequence of Chebyshev parameters, taken in a definite order. To find λ_1 we shall use the so-called T -sequence of parameters, while to find ϕ_1 we use the U -sequence. These infinite sequences were described and studied in [2]; they provide convergence of the iterations which is optimal in a definite class of errors, for a certain set of numbers of iterations

$k_i \rightarrow \infty$, and ensure computational stability with respect to rounding errors. In [3, 4], a cyclical method with Chebyshev parameters was described, for finding the eigenfunction. To find λ_1 and λ_2 , we extend to the case of variable displacements the well-known Bernoulli method [3, 5, 6]; this enables the eigenvalues of interest to be found more rapidly. This method is known to provide substantial acceleration of the convergence of $\lambda_1^{(k)}$ to λ_1 when λ_1 and λ_2 are close together.

Before proceeding to find the eigenfunction, we turn our attention to the fact that the class of initial errors is transformed after performing the iterations for finding λ_1 . Before proceeding to "uniform" suppression of the error components, we suppress the largest components by means of four intermediate iterations.

We consider the following iterative method for finding λ_1, ϕ_1 : given the initial approximation

$$x^0 = \sum_{n=1}^{\infty} C_n^0 \varphi_n,$$

for which we assume that $l(x^0) = 1$, $C_1^0 \neq 0$, and assuming that the error belongs to the class

$$|C_n^0| \leq C_0, \quad n=2, 3, \dots, \quad (6)$$

where, here and below, $C_i > 0$ are constants, we find the approximations

$$x^h = \sum_{n=1}^{\infty} C_n^h \varphi_n$$

from the expressions

$$\tilde{x}^{h+1} = A x^h - \beta_{h+1} x^h, \quad x^{h+1} = \tilde{x}^{h+1} / l(\tilde{x}^{h+1}), \quad (7)$$

where

$$\beta_h = [M + m + (M - m) \cos(\pi \omega_h)] / 2; \quad (8)$$

$\{\omega_h, h=1, 2, \dots, \infty\}$ is an infinite sequence, $\omega_h \in (0, 1)$; M, m are parameters, which we have at our disposal at each step of finding λ_1, ϕ_1 .

Stage 1. We first find λ_1, λ_2 . For this, we put $m = 0, M = a$, while as $\{\omega_h, h=1, 2, \dots, \infty\}$ we take a T -sequence [2]. We denote by $T(N, p)$ the T -sequence for which $\{\cos(\pi \omega_k), k=1, 2, \dots, Np^n\}$ are the same as the roots $T_{Np^n}(x)$ of a Chebyshev polynomial of the first kind. To find λ_1, λ_2 , we use the $T(2,3)$ -sequence in (8). Then, $0 < \beta_h < a$ and

$$\begin{aligned} \tilde{x}^h &= (\lambda_1 - \beta_h) \left[\varphi_1 + \sum_{n=2}^{\infty} \psi_h(\lambda_n) \frac{C_n^0}{C_1^0} \varphi_n \right] \left[l_1 + \sum_{n=2}^{\infty} \psi_{h-1}(\lambda_n) \frac{C_n^0}{C_1^0} l_n \right]^{-1}, \\ x^h &= \left[\varphi_1 + \sum_{n=2}^{\infty} \psi_h(\lambda_n) \frac{C_n^0}{C_1^0} \varphi_n \right] \left[l_1 + \sum_{n=2}^{\infty} \psi_h(\lambda_n) \frac{C_n^0}{C_1^0} l_n \right]^{-1} \end{aligned} \quad (9)$$

$$\gamma_k = l(\hat{x}^k) = (\lambda_1 - \beta_k) \left[1 + \sum_{n=2}^{\infty} \psi_k(\lambda_n) \frac{C_n^0 l_n}{C_1^0 l_1} \right] \left[1 + \sum_{n=2}^{\infty} \psi_{k-1}(\lambda_n) \frac{C_n^0 l_n}{C_1^0 l_1} \right]^{-1}, \quad (10)$$

where

$$\psi_k(\lambda) = P_k(\lambda) / P_k(\lambda_1), \quad P_k(\lambda) = \prod_{i=1}^k (\lambda - \beta_i).$$

Let $\theta_n = 2\lambda_n/a - 1$, then, for $\lambda_n \geq a$, we have

$$0 \leq \psi_k(\lambda_n) \leq \frac{T_{\tilde{k}}(\theta_n)}{T_{\tilde{k}}(\theta_1)} \left(\frac{\lambda_n - a/2}{\lambda_1 - a/2} \right)^{k-\tilde{k}} \leq \left(\frac{\lambda_n - a/2}{\lambda_1 - a/2} \right)^k, \quad (11)$$

where $\tilde{k} = \max j, 2 \cdot 3^j \leq k$, while for $\lambda_n \in [0, a]$ we have [2]

$$|\psi_k(\lambda_n)| \leq \alpha(\Delta) / T_k(\theta_1). \quad (12)$$

where $\Delta = \lambda_1/a - 1$, and $\alpha(\Delta) = 1$ for $k = \tilde{k}$ or for $\lambda_n > 2a - \lambda_1$ and $\alpha(\Delta) = C\Delta^{-1}$, $C > 0$, otherwise.

It follows from (9)–(12) that the coefficients of ϕ_n, l_n in (9), (10) for $n \geq 2$ decreases in modulus more rapidly than in the method of simple iteration as $k \rightarrow \infty$, while for $k = \tilde{k}$ class-(6)-optimal suppression occurs of the coefficients C_n^0 for which $\lambda_n \in [0, a]$.

Assume that the values of $\beta_k, \beta_{k+1}, \beta_{k+2}, \gamma_k, \gamma_{k+1}, \gamma_{k+2}$ are known. Then, putting $C_1 = C_1^{k-1} l_1, C_2 = C_2^{k-1} l_2$ and retaining only the first two terms in the sums (10) (on the assumption that the terms with $n > 2$ are small for sufficiently large k , see (11), (12)), we obtain the system of four equations:

$$\begin{aligned} C_1 + C_2 &= D_k, \\ \sum_{n=1}^2 C_n \prod_{i=0}^m (\lambda_n - \beta_{k+i}) &= D_k \prod_{i=0}^m \gamma_{k+i}, \quad m=0, 1, 2, \end{aligned}$$

where D_k is a non-zero constant.

This is a generalized moment system. If

$$\begin{aligned} \sigma_k &= \frac{\gamma_{k+2} - \gamma_{k+1} + \beta_{k+2} - \beta_{k+1}}{\gamma_{k+1} - \gamma_k + \beta_{k+1} - \beta_k} \gamma_{k+1}, \\ q_k &= \gamma_k \sigma_k, \quad p_k = \gamma_{k+1} + \beta_{k+1} - \beta_k + \sigma_k, \\ t_1 &= p_k/2 + (p_k^2/4 - q_k)^{1/2}, \quad t_2 = q_k/t_1, \end{aligned} \quad (13)$$

where t_1, t_2 are the roots of the equation $t^2 - p_k t + q_k = 0$, then

$$\lambda_1 = t_1 + \beta_k, \quad \lambda_2 = t_2 + \beta_k. \quad (14)$$

This stage of the iterations, using the T -sequence, is continued on the basis of expressions (7), (8), (13), (14) until we obtain stable values of λ_1, λ_2 . Assume that this occurs for $k = k_1$, and assume that the eigenfunction determination accuracy criterion for terminating the iterations has not yet been satisfied.

If $\lambda_2 \leq a$, we continue the iterative process (7), (8) with the T sequence until the eigenfunction is obtained to the required accuracy.

Stage 2. If $\lambda_2 > a$, the class of errors (6) is transformed after k_1 iterations to the class

$$|C_n^{k_1}| \leq C_1 |\psi_{k_1}(\lambda_n)|. \quad (15)$$

If $\lambda = \lambda_2$, the function $|\psi_{k_1}(\lambda)|$ with large k_1 has a sharp peak; this means that the error contains at this instant relatively large components $C_n^{k_1} \varphi_n$ for the first $n \geq 2$. We smooth the resultant non-uniformity in the error by four iterations. For this, we first put

$$\beta_{k_1+1} = \lambda_2, \quad \beta_{k_1+2} = 0. \quad (16)$$

this choice of parameters eliminates the maximum of the transfer function for $\lambda = \lambda_2$, and smoothes the rounding errors appearing in the iterations in the components with large n , as a result of the subsequent use of the U sequence. Let

$$p_2(\lambda) = (\lambda - \beta_{k_1+3})(\lambda - \beta_{k_1+4}), \quad \Phi(\lambda) = |\psi_{k_1}(\lambda)(\lambda(\lambda_2 - \lambda))^{1/2}|.$$

The ideal choice of $\beta_{k_1+3}, \beta_{k_1+4}$, would be that for which

$$\inf_{\beta_{k_1+3}, \beta_{k_1+4}} \max_{0 \leq \lambda \leq \lambda_2} |p_2(\lambda) \Phi(\lambda)|.$$

is reached. This choice presents a serious problem, however. We shall simplify it in two ways. First, we replace $\Phi(\lambda)$ by a simpler function, retaining the characteristic properties of $\Phi(\lambda)$:

$$\begin{aligned} \Phi^2(\lambda) &\sim C_2 \lambda (\lambda_2 - \lambda) (1 + T_{2k_1}(2\lambda/a - 1)) / T_{k_1}^2(\theta_1) \\ &\sim C_2 (\lambda(\lambda_2 - \lambda))^{1/2} \Phi_1(\lambda/\lambda_2) / T_{k_1}^2(\theta_1), \end{aligned}$$

where

$$\begin{aligned} \Phi_1(x) &= (1-x)^{1/2} \left(x^{1/2} + \tilde{\theta}(x-b) A \left(\frac{x-b}{1-b} \right)^{2k_1+1/2} \right), \\ x &= \lambda/\lambda_2, \quad \tilde{\theta}(x) = 0 \text{ for } x < 0, \quad \tilde{\theta}(x) = 1 \text{ for } x \geq 0, \\ A &= T_{2k_1}(\theta_2), \quad b = (1-s)/(1-s/4k_1), \quad s = a/\lambda_2. \end{aligned}$$

The quantities A and b are chosen in such a way that, with $\lambda = \lambda_2, i=0, 1$

$$\frac{d^i}{dx^i} \lambda^{1/2} T_{2k_1}(2\lambda/a - 1) = A \frac{d^i}{dx^i} \left(\frac{\lambda/\lambda_2 - b}{1-b} \right)^{2k_1+1/2}.$$

Then, on the basis of the results of [7], we define β_{k_1+3} , β_{k_1+4} as the solution of the following problem: to find

$$\min_{\beta_{k_1+3}, \beta_{k_1+4}} \int_0^{\lambda_2} p_2^2(\lambda) \Phi_1(\lambda/\lambda_2) d\lambda. \quad (17)$$

Problem (17) can be solved in explicit form:

$$\beta_{k_1+3}, \beta_{k_1+4} = \lambda_2 (b + (1-b)(p \pm \bar{m})), \quad (18)$$

where $p = (t_1 t_4 - t_2 t_3) / 2(t_1 t_3 - t_2^2)$, $\bar{m} = (p^2 - (t_2 t_4 - t_3^2) / (t_1 t_3 - t_2^2))^{1/2}$, $t_i = l_i + m_i$, $l_1 = 1$, $m_1 = (\theta_2 - (\theta_2^2 - 1)^{1/2})^{2k_1} (4k_1 - 3) [(k_1 + 1)(\pi/8)(1-b)^{-3}]^{1/2}$, $l_i = l_{i-1} (4k_1 + 2i - 1) / (4k_1 + 2i + 2)$, $i = 2, 3, 4$, $m_2 = m_1(1 - 2y)$, $m_3 = m_1(5y^2 - 4y + 1)$, $m_4 = m_1(15y^2 - 14y^3 - 6y + 1)$, $y = [4(1-b)]^{-1}$.

If the required accuracy of finding φ_1 has not been achieved at this instant, the iterative process (7), (8) can be further continued in two ways.

Stage 3'. We put $M = \lambda_2$, $m = 0$ in Eq. (8) and again, starting with ω_1 , use the T sequence.

Stage 3''. Detailed account can be taken of the information obtained from the previous iterations, and the parameters ω_k more accurately chosen on this basis. Let

$$h = \max_{0 \leq \lambda \leq \lambda_2} \Phi_1(\lambda/\lambda_2) / T_{k_1}(\theta_1).$$

We know that the function $\psi_{k_1+4}(\lambda)$ vanishes for $\lambda = 0, \lambda_2$. Then, in the class of errors

$$|C_n^{k_1+4}| \leq C_2 h (\lambda_n (\lambda_2 - \lambda_n))^{1/2}, \quad n \geq 2, \quad (19)$$

it is advisable to use the following parameters in the method (7), (8):

$$M = \lambda_2, \quad m = 0, \quad \omega_{k_1+4+p} = \alpha_p,$$

where $\{\alpha_p, p = 1, 2, \dots, \infty\}$ is any U sequence [2]. By $U(N)$ we denote the U sequence for which $\{\cos(\pi\alpha_p), p = 1, 2, \dots, 2^m(N+1)-1\}$ are the same as the roots of the Chebyshev polynomial of the 2nd kind of degree $2^m(N+1) - 1$. For clarity, we shall use the $U(1)$ -sequence in (8).

With this choice of ω_k , optimal suppression of the coefficients $C_n^{k_1+4+p}$ is achieved, for all $p = 2^n - 1$, $n = 1, 2, \dots$ in the class (19), and the stability of the iterative process with respect to rounding errors is preserved.

Notes. 1. In our construction of the iterative method (7), (8), (13), (14), we took account of certain unfavourable situations which may occur when realizing it numerically (e.g., $\lambda_2 \approx \lambda_1$, $a \ll \lambda_1$, etc.). When the actual situation is favourable (the iterative process converges well), the expressions obtained cannot slow down the convergence of the iterations.

2. For the boundary value problem (1)–(3), Ax^k is evaluated by an iterative process, which is best started with the values $(\varphi_1^h, \dots, \varphi_n^h)$, obtained at the previous exterior iteration, while taking account of the last normalization.

3. If, during the last stage of the iterations, the quantity $\lambda_2^{(k)}$ (see (14)) again takes the stable value $\bar{\lambda}_2^{(k)}$ (i.e., at this instant the main part of the error is concentrated in the eigenfunctions with eigenvalues in the neighbourhood of $\bar{\lambda}_2^{(k)}$), then we put

$$\beta_{k+1} = \bar{\lambda}_2^{(k)} \quad (20)$$

and we continue the iterations with the U sequence.

2. Construction of the T and U sequences

The $T(2, 3)$ sequence. We shall first define the sequence of permutations κ_3^n by the following expressions [8]: $\kappa_1 = (1)$. If we know the permutation

$$\kappa_{3^{n-1}} = (j_1, \dots, j_{3^{n-1}}), \quad (21)$$

where $1 \leq j_k \leq 3^{n-1}$, then we define the permutation κ_3^n by the expression

$$\kappa_3^n = (j_1, 2 \cdot 3^{n-1} + j_1, 2 \cdot 3^{n-1} + 1 - j_1, \dots, j_k, 2 \cdot 3^{n-1} + j_k, 2 \cdot 3^{n-1} + 1 - j_k, \dots).$$

We put

then,

$$t_1 = 2^{-1/2}, \quad t_2 = -t_1. \quad (22)$$

Assume that a segment of the sequence $\{t_k, k=1, 2, \dots, 2 \cdot 3^{n-1}\}$ has been constructed. Given the permutation κ_3^{n-1} (see (21)), we construct the segment $\{t_k, k=2 \cdot 3^{n-1} + 1, \dots, 2 \cdot 3^n\}$ from the expressions

$$t_{2 \cdot 3^{n-1} + 4l + 1} = \sin \frac{2(j_{l+1} + [j_{l+1}/2]) - 1}{4 \cdot 3^n} \pi, \quad t_{2 \cdot 3^{n-1} + 4l + 2} = -t_{2 \cdot 3^{n-1} + 4l + 1}, \quad (23)$$

$$t_{2 \cdot 3^{n-1} + 4l + 3} = (1 - t_{2 \cdot 3^{n-1} + 4l + 2})^{1/2}, \quad t_{2 \cdot 3^{n-1} + 4l + 4} = -t_{2 \cdot 3^{n-1} + 4l + 3},$$

$$l = 0, 1, \dots, 3^{n-1} - 1.$$

After this, we form κ_3^n and evaluate $\{t_k, k=2 \cdot 3^n + 1, \dots, 2 \cdot 3^{n+1}\}$, etc.

For each $k = 2 \times 3^n$, this sequence gives the class-(6)-optimal convergence with $\lambda_n \in [0, a]$.

The $U(1)$ sequence. We first define the sequence of permutations κ_2^n by the following expressions [9]. If we know the permutation $\kappa_{2^{n-2}} = (j_1, \dots, j_{2^{n-2}})$, where $1 \leq j_k \leq 2^{n-2}$, then the permutation κ_{2^n-1} is found from the expression

$$\kappa_{2^n-1} = (j_1, 2^{n-1} + 1 - j_1, \dots, j_k, 2^{n-1} + 1 - j_k, \dots).$$

We put $u_k = \cos(\pi \alpha_k)$, then,

$$u_1 = 0, \quad u_2 = 2^{-1/2}, \quad u_3 = -u_2. \quad (24)$$

Let the sequence $\{u_k, k=1, 2, \dots, 2^n-1\}$ be already constructed. Knowing the permutation κ_{2^n-2} , we construct the segment $\{u_k, k=2^n, \dots, 2^{n+1}-1\}$ from the expressions

$$\begin{aligned} u_{2^n+4l} &= \sin \frac{2j_{l+1}-1}{2^{n+1}} \pi, & u_{2^n+4l+1} &= -u_{2^n+4l}, \\ u_{2^n+4l+2} &= (1-u_{2^n+4l+1})^{1/2}, & u_{2^n+4l+3} &= -u_{2^n+4l+2}, \\ l &= 0, 1, \dots, 2^{n-2}-1. \end{aligned} \quad (25)$$

After this, we form κ_{2^n-1} and evaluate $\{u_k, k=2^{n+1}, \dots, 2^{n+2}-1\}$ etc.

For each $k = 2^n - 1$, this sequence gives the optimal convergence in the class (19).

3. Program

The ALGOL program realizing our method is as follows. The program is written in the form of a block. There are the following correspondences between the program identifiers and the notation of the present paper:

$$\begin{aligned} bi &= \beta_{k+i-3}, & gi &= \gamma_{k+i-3}, & i &= 1, 2, 3, & l0 &= \lambda_1^{(k-1)}, & l1 &= \lambda_1^{(k)} \\ l2 &= \lambda_2^{(k)}, & kp &= \kappa. \end{aligned}$$

The program uses access to the procedure ITER ($b3, g3, d$), which, in accordance with the values $b3, x^{k-1}$ evaluates x^k from expressions (7), computes the quantities $g3$ and $d = \|x^k - x^{k-1}\|$ and sends $x^k \rightarrow x^{k-1}$. In the program, $k2 = 0, 1, 2, 3$, depending on whether the parameters are computed from expressions (22), (23), or (16), (18), (24), or (25), or (20) respectively. If $d < \text{eps}$ (this condition can easily be replaced by another), the computation stops. For simplicity it is assumed that the blocks, connected with the operator A and x^k, x^{k-1} , and also the quantity eps are described and specified in some external block.

```
begin integer k, k1, k2, i, j, n; real b1, b2, b3, g1, g2, g3, l0, l1, l2, l3, cs, a, d, m,
M, p, p1, p2, y; array kp; label A1, A2, A3, A4;
k:=0; M:=l1:=a; l2:=g2:=g3:=b2:=b3:=a/2; p:=0.261799387799; k1:=4;
k2:=1;
A1: l0:=l1; g1:=g2; g2:=g3; b1:=b2; l3:=l2; b2:=b3; k:=k+1; k1:=k1+1,
if k2=1 then begin
if k1=5 then begin cs:=0.707106781187; go to A3 end;
if k1=6 then begin n:=i:=kp[1]:=1; k1:=0;
if k=2 then begin k2:=0; g1:=g2; b1:=b2 end else k2:=2; go to A2 end;
if k1=0 then b3:=M; if k1=1 then b3:=0;
if k1=2 then b3:=cs; if k1=3 then b3:=p2;
if k1=4 then b3:=M/2; go to A4 end;
if k2=3 then begin b3:=l2; k2:=2; l3:=0; k1:=k1-1; go to A4 end;
if k1=5 then begin k1:=1; i:=i+1 end;
if k1=3 then begin cs:=sqrt(1-cs^2); go to A3 end;
if k1=1 then begin if k2=0 then begin
j:=kp[i]+entier(kp[i]/2); cs:=sin(p*(2*j-1)/n) end else cs:=sin(p*(2*j-1)/n); go to A3 end;
comment iterations with parameters (22), (23), (26);
if i=n^k1=4 then begin
if k2=0 then begin i:=3*n-2;
for j=n step -1 until 1 do begin
kp[i]:=kp[j]; kp[i+1]:=2*n+kp[j]; kp[i+2]:=2*n+1-kp[j]; i:=i-3
```

```

    end n:=3×n end else begin i:=2×n-1;
    for j=n step -1 until 1 do begin
        kp[i]:=kp[j]; kp[i+1]:=2×n-1-kp[j]; i:=i-2 end; n:=2×n end; i:=0 end;
    comment new κ is obtained;
A2: cs:=-cs;
A3: b3:=M×(1-cs)/2;
A4: procedure ITER (b3, d, g3);
    m:=b3+g3; y:=g2-g1+b2-b1;
    if abs(y)>10↑(-9) then begin
        y:=g2×(m-g2-b2)/y; l1:=(g2+b2-b1+y)/2; p1:=g1×y; y:=l1↑2-p1;
    if y<0 then begin l1:=m; l2:=0; go to A1 end; l1:=l1+sqrt(y); l2:=p1/l1+b1;
        l1:=l1+b1;
    if k2≠0 ∧ abs(l1-m)>abs(l2-m) then begin y:=l1; l1:=l2; l2:=y end end
    else begin l1:=m; l2:=0 end;
    comment new λ1, λ2 are obtained;
    y:=abs(l1-l0); if y>eps then go to A1;
    if k2=0 then begin
        if a<l2 ∧ abs((l2-l3)/l1)<5×10↑(-3)
        then begin comment iterations with T-sequence are terminated;
            k2:=1; p:=3×p; k1:=4×k+4; l0:=sqrt(1-a/l2); l0:=(1-l0)/(1+l0/(k1-4));
            y:=1/(4×(1-l0));
            if a=0 then m:=0 else begin
                m:=2×l2/a-1; m:=(m-sqrt(m↑2-1))↑(2×k)×(k1-1)×2×y×sqrt(2×
                p×k1×y) end; cs:=1+m; l3:=(k1-1)/(k1+2); p1:=m×(1-2×y)+l3;
                l3:=l3×(k1+1)/(k1+4); p2:=m×(y×(5×y-4)+1)+l3; l3:=m×(y×
                (y×(15-14×y)-6)+1+l3×(k1+3)/(k1+6)); y:=cs×p2-p1↑2; m:=
                cs×l3-p1×p2; M:=p1×l3-p2↑2; p2:=m/(2×y); k1:=-1;
                m:=sqrt(p2↑2-M/y); M:=l2; cs:=l2×(l0+(1-l0)×(p2+m)); p2:=l2×
                (l0+(1-l0)×(p2-m)); end end;
            if k2=2 ∧ abs((l2-l3)/l1)<5×10↑(-3) ∧ l1>l2>0 ∧ k1<3 then k2:=3;
            if d>eps then go to A1 end;

```

Computations of the problem (1)–(3), performed by S. A. Frolov, Yu. A. Vlasov, and S. I. Konyaev, showed that the present method offers a considerable advantage in convergence rate over the power method.

Translated by D. E. Brown

REFERENCES

1. MARCHUK, G. I., *Methods of nuclear reactor design* (Metody rasheta yadernykh reaktorov), Atomizdat, Moscow, 1961.
2. LEBEDEV, V. I., and FINOGENOV, S. A., On the study of ordered Chebyshev parameters in iterative methods, *Zh. vychisl. Mat. mat. Fiz.*, **16**, No. 4, 895–907, 1976.
3. FLANDERS, D. A., and SHORTLEY, G., Numerical determination of fundamental modes, *J. Appl. Phys.*, **21**, 1326–1332, 1950.
4. LEBEDEV, V. I., and FINOGENOV, S. A., On an algorithm for choosing the parameters in Chebyshev cyclical methods, in: *Computational methods of linear algebra* (Vychisl. metody lineinoi algebr), VTs SO Akad Nauk SSSR, 21–27, Novosibirsk, 1972.
5. AITKEN, A., Studies in practical mathematics, II, The evaluation of the latent roots and latent vectors of a matrix, *Proc. Roy. Soc. Edinburgh Sec. A*, **57**, 269–304, 1936.
6. FADDEEV, D. K., and FADDEEVA, V. N., *Computational methods of linear algebra* (Vychislitel'nye metody lineinoi algebr), Fizmatgiz, Moscow, 1963.

7. BERNSTEIN, S. N., On polynomials orthogonal in a finite interval, *Sobr. soch.* Vol. 2, 7-106, Izd-vo Akad. Nauk SSSR, 1954.
8. LEBEDEV, V. I., and FINOGENOV, S. A., Solution of the problem of ordered parameters in Chebyshev iterative methods, *Zh. vychisl. Mat. mat. Fiz.*, 13, No. 1, 18-33, 1973.
9. LEBEDEV, V. I., and FINOGENOV, S. A., On the order of choosing the iterative parameters in a Chebyshev cyclical iterative method, *Zh. vychisl. Mat. mat. Fiz.*, 11, No. 2, 425-438, 1971.

A VARIATIONAL-DIFFERENCE METHOD FOR SOLVING TWO-DIMENSIONAL LINEAR PARABOLIC EQUATIONS*

Yu. R. AKOPYAN and L. A. OGANESYAN

Leningrad

(Received 26 May 1975; revised 4 November 1975)

IMPLICIT variational-difference schemes are constructed for the first and third initial-boundary value problems for linear parabolic equations in a two-dimensional domain with a smooth boundary. It is assumed that the coefficients of the equation are sufficiently smooth and that the right-hand side belongs to the space L_2 . Order-wise exact estimates are obtained for the convergence rate of the schemes in the norm of the energy space; the estimates are equal to the diameter of the set of solutions of the differential equations in this space.

Below we construct implicit variational-difference schemes (v.d.s) for the first and third initial-boundary value problems for two-dimensional linear parabolic equations with a right-hand side belonging to $L_2(\Omega \times (0, T))$, in which the time derivative is replaced by the backwards difference ratio. A discussion of v.d.s. for such problems can be found in [1-4]; there, convergence rate estimates are obtained for the schemes under various assumptions about the smoothness of the right-hand side. In the present paper no assumption is made about the smoothness of the right-hand side, and it is found that the schemes are optimal in a sense for $\tau = O(h^2)$, where h is the mesh step, and τ is the time step; the convergence rate estimate is of the $O(h)$.

1. Notation, and construction of the v.d.s.

1. In a cylindrical domain $Q = \Omega \times (0, T)$ with lateral surface Γ , where Ω is a bounded simply connected domain of space R_2 of points $(x_1, x_2) = (x, y)$ with boundary $S \in C^2$, we consider the equation

$$Lu = \frac{\partial u}{\partial t} - \sum_{i,j=1}^2 \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial u}{\partial x_j} \right) + \sum_{i=1}^2 b_i \frac{\partial u}{\partial x_i} + au = f \quad (1.1)$$

with the initial condition

$$u|_{t=0} = 0 \quad (1.2)$$

and the boundary condition

$$u|_{\Gamma} = 0, \quad (1.3)$$

**Zh. vychisl. Mat. mat. Fiz.*, 17, 1, 109-118, 1977.

$$\text{or} \quad \left[\sum_{i,j=1}^2 a_{ij} \frac{\partial u}{\partial x_i} \cos(\nu, x_i) + \sigma u \right] \Big|_{\Gamma} = 0, \quad (1.4)$$

where ν is the outward normal to S .

We denote the norm in Sobolev–Slobodetskii space $W_2^{m,n}(Q)$ [5] by $\|\cdot\|_{m,n,Q}$, the norm in Sobolev space $W_2^m(\Omega)$ by $\|\cdot\|_{m,Q}$, and the norm in $L_2(Q)$ by $\|\cdot\|_{0,Q}$. We denote by $V_2^{1,0}(Q)$ the space consisting of all the elements of $W_2^{1,0}(Q)$, having the finite norm

$$|u|_Q = \sup_{0 \leq t \leq T} \|u(t)\|_{0,Q} + \|\nabla u\|_{0,Q},$$

where

$$|\nabla u| = \left(\left| \frac{\partial u}{\partial x} \right|^2 + \left| \frac{\partial u}{\partial y} \right|^2 \right)^{1/2}.$$

We assume that

$$\begin{aligned} a_{ij} &\in C^{2,1}(Q), \quad b_i \in C^{1,0}(Q), \quad a \in C(Q), \quad \sigma \in C^1(\Gamma), \quad f \in L_2(Q), \\ 0 < \mu_0 \sum_{i=1}^2 \xi_i^2 &\leq \sum_{i,j=1}^2 a_{ij} \xi_i \xi_j \leq \mu_1 \sum_{i=1}^2 \xi_i^2, \quad \mu_0, \mu_1 = \text{const.} \end{aligned} \quad (1.5)$$

Then, an initial-boundary value problem has a unique solution $u \in W_2^{2,1}(Q)$ and we have the estimate [6]

$$\|u\|_{2,1,Q} \leq C \|f\|_{0,Q}. \quad (1.6)$$

Throughout, the letter C with or without subscripts denotes a positive constant, regardless of any factors that may be present.

2. We specify a positive parameter h , which we call the mesh step.

Let us construct the mesh domain Ω^h for the third initial-boundary value problem. For this, we superimpose a square mesh of step h on the domain Ω , in such a way that the mesh lines are parallel to the coordinate axes. A regular rectangular mesh may also be taken, in which the lengths of sides of the cell are of order h . We divide the mesh cells into triangles by a diagonal at an angle $\pi/4$ to the x_1 axis. As Ω^h we take the least union of triangles containing $\bar{\Omega}$.

For problem (1.1)–(1.3) with respect to the domain Ω , we define the mesh domain Ω^h with boundary S^h , which satisfies the following conditions [7, 8]: 1) the domain Ω^h , bounded by the step-line S^h , lies in the domain Ω ; 2) between points of the step-line S^h and S we establish a one-to-one correspondence with the aid of the normals to S ; 3) the lengths of the sections of the step-line S^h are bounded from below by lh ; 4) the distances from points of S^h to S do not exceed δh^2 . We then divide the domain Ω^h into triangles [7], the lengths of sides of which lie in the range $[l_1 h, l_2 h]$, and their areas in the range $[s_1 h^2, s_2 h^2]$. Here, the positive constants $l, \delta, l_1, l_2, s_1, s_2$ are independent of h . The choice of these constants is determined by the properties of the curve S and the algorithm for constructing Ω^h .

Henceforth, unless stipulated otherwise, Ω^h means the mesh domain, constructed for the first or third initial-boundary value problem.

The set of vertices and sides of the triangles of the triangulation form a mesh, and the vertices of the triangles will be called the mesh base-points. We shall assume that all the base-points are numbered in a certain order. We denote by (m) the m -th base-point (x_m, y_m) . We introduce the following notation: R^h is the set of base-points belonging to Ω^h ; r^h is the set of base-points belonging to S^h ; and \bar{R}^h is the set of base-points belonging to $\bar{\Omega}^h$.

We put $Q^h = \Omega^h \times (0, T)$. We divide the interval $[0, T]$ into equal parts with the step τ , $t_n = n\tau$, $n = 0, 1, \dots, N$.

For each base-point $(m) \in \bar{R}^h$ we define the function $\phi_m(x, y)$, which is equal to unity at the base-point (m) , and to zero at the other base-points, while it is interpolated piecewise linearly in $\bar{\Omega}^h$. Outside $\bar{\Omega}^h$ the function is identically zero. We put

$$\varphi_{mn}(x, y, t) = \begin{cases} \phi_m(x, y), & \text{if } t \in (t_{n-1}, t_n], \\ 0, & \text{if } t \in (t_{n-1}, t_n], \end{cases}$$

where $(m) \in \bar{R}^h$, $n = 1, 2, \dots, N$.

Let $\underline{v} = \{v_m\}$ and $\underline{w} = \{w_{mn}\}$ be the mesh functions, specified at the points (x_m, y_m) and (x_m, y_m, t_n) respectively. We introduce the notation:

$$\bar{v}(x, y) = \sum_{(m) \in \bar{R}^h} v_m \phi_m(x, y),$$

$$\bar{w}(x, y, t) = \sum_{n=1}^N \sum_{(m) \in \bar{R}^h} w_{mn} \varphi_{mn}(x, y, t), \quad (1.7)$$

$$\hat{w}(x, y, t) = \sum_{n=1}^N \sum_{(m) \in \bar{R}^h} w_{mn} \varphi_{mn}(x, y, t). \quad (1.8)$$

The set of functions of the type (1.7) will be denoted by $H_{h\tau}$, and of the type (1.8), by $\hat{H}_{h\tau}$.

The arguments x, y , and also t , will sometimes be omitted when writing a function.

3. We put

$$\mathcal{L}_1(u, \Phi) = \sum_{i,j=1}^2 \int_Q a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial \Phi}{\partial x_i} dQ + \sum_{i=1}^2 \int_Q b_i \frac{\partial u}{\partial x_i} \Phi dQ + \int_Q au \Phi dQ,$$

$$\mathcal{L}_3(u, \Phi) = \mathcal{L}_1(u, \Phi) + \int_{\Gamma} \sigma u \Phi d\gamma,$$

where, here and below, $dQ = dx dy dt$.

For the first initial-boundary value problem, as the approximate solution we take the function $\hat{v} \in \hat{H}_{h\tau}$, which satisfies the integral identity

$$\tau \sum_{n=1}^N \int_Q (\hat{v}(t_n))_i \hat{\varphi}(t_n) d\Omega + \mathcal{L}_1(\hat{v}, \hat{\varphi}) = \int_Q f \hat{\varphi} dQ \quad (1.9)$$

for arbitrary $\hat{\varphi} \in \hat{H}_{h\tau}$, where $(\hat{v}(t_n))_i$ is $\tau^{-1}(\hat{v}(t_n) - \hat{v}(t_{n-1}))$. Here and below $d\Omega = dx dy$.

As the approximate solution of the third initial boundary value problem we take a function $\tilde{v} \in H_{ht}$, which satisfies the integral identity

$$\tau \sum_{n=1}^N \int_{\Omega} (\tilde{v}(t_n))_{,\tau} \tilde{\varphi}(t_n) d\Omega + \mathcal{L}_s(\tilde{v}, \tilde{\varphi}) = \int_{\Omega} f \tilde{\varphi} dQ \quad (1.10)$$

for arbitrary $\tilde{\varphi} \in H_{ht}$.

2. Approximation theorems

Denote by u^h the Steklov average of the function u :

$$u^h(x, y, t) = \frac{1}{4h^2} \int_{x-h}^{x+h} \int_{y-h}^{y+h} u(\xi, \eta, t) d\xi d\eta.$$

Theorem 1

If $u \in W_2^{2,1}(Q)$ and is continued into the space R_3 of points (x, y, t) while retaining its class and norm [5], then we have

$$\|u - \tilde{u}^h\|_{Q^h} \leq C(h + \tau^{1/2} + \tau h^{-1}) \|u\|_{2,1,Q}, \quad (2.1)$$

$$\|u - \tilde{u}^h\|_{0,Q^h} \leq C(h^2 + \tau) \|u\|_{2,1,Q}. \quad (2.2)$$

Proof. We have the inequality

$$\| |\nabla(u - \tilde{u}^h)| \|_{0,Q^h} \leq \| |\nabla(u - u^h)| \|_{0,Q^h} + \| |\nabla(u^h - \tilde{u}^h)| \|_{0,Q^h}. \quad (2.3)$$

Further,

$$\begin{aligned} \| |\nabla(u^h - \tilde{u}^h)| \|_{0,Q^h} &\leq \left\{ \int_0^T \| |\nabla(u^h(t) - \overline{u}^h(t)) | \|_{0,Q^h}^2 dt \right\}^{1/2} \\ &+ \left\{ \int_0^T \| |\nabla(\overline{u}^h(t) - \tilde{u}^h(t)) | \|_{0,Q^h}^2 dt \right\}^{1/2}. \end{aligned} \quad (2.4)$$

Noting that $\tilde{u}^h(t) = \overline{u}^h(t_n)$ for $t \in (t_{n-1}, t_n]$, we get

$$\left\{ \int_0^T \| |\nabla(\overline{u}^h(t) - \tilde{u}^h(t)) | \|_{0,Q^h}^2 dt \right\}^{1/2} \leq C\tau h^{-1} \|u\|_{2,1,Q}. \quad (2.5)$$

The following estimates hold (see e.g., [8]):

$$\| |\nabla(u - u^h)| \|_{0,Q^h} \leq Ch \|u\|_{2,1,Q}, \quad (2.6)$$

$$\left\{ \int_0^T \| |\nabla(u^h(t) - \overline{u}^h(t)) | \|_{0,Q^h}^2 dt \right\}^{1/2} \leq Ch \|u\|_{2,1,Q}. \quad (2.7)$$

It follows from (2.3)–(2.7) that

$$\| |\nabla(u - \tilde{u}^h)| \|_{0, Q^h} \leq C(h + \tau h^{-1}) \|u\|_{2, 1, Q}. \quad (2.8)$$

We can show in just the same way that

$$\sup_{0 \leq t \leq T} \|u(t) - \tilde{u}^h(t)\|_{0, Q^h} \leq C\tau^{1/2} \|u\|_{2, 1, Q} + Ch \sup_{0 \leq t \leq T} \|u(t)\|_{1, Q^h},$$

whence, using the inequality [5]

$$\sup_{0 \leq t \leq T} \|u(t)\|_{1, Q^h} \leq C \|u\|_{2, 1, Q},$$

we obtain

$$\sup_{0 \leq t \leq T} \|u(t) - \tilde{u}^h(t)\|_{0, Q^h} \leq C(h + \tau^{1/2}) \|u\|_{2, 1, Q}. \quad (2.9)$$

From (2.8) and (2.9) we get inequality (2.1).

The proof of inequality (2.2) is essentially the same as the proof of (2.1) and may therefore be omitted. The theorem is proved.

Let Ω^h be the mesh domain constructed for the first initial-boundary value problem. We have:

Theorem 2

Let the function $u \in W_2^{2,1}(Q)$, $u|_{\Gamma} = 0$ and be continued into R_3 while retaining its class and norm [5]. Then, we have the inequalities

$$|u - \hat{u}^h|_Q \leq C(h + \tau^{1/2} + \tau h^{-1}) \|u\|_{2, 1, Q}, \quad (2.10)$$

$$\|u - \hat{u}^h\|_{0, Q} \leq C(h^2 + \tau) \|u\|_{2, 1, Q}. \quad (2.11)$$

Proof. Since $\hat{u}^h = 0$ in $Q \setminus Q^h$, we have

$$|u - \hat{u}^h|_Q \leq |u|_{Q \setminus Q^h} + |u - \tilde{u}^h|_{Q^h} + |\tilde{u}^h - \hat{u}^h|_{Q^h}. \quad (2.12)$$

Noting that the width of the strip $Q \setminus Q^h$ is of order h^2 , we obtain [8, 9]

$$|u|_{Q \setminus Q^h} \leq Ch \|u\|_{2, 1, Q}. \quad (2.13)$$

The second term on the right-hand side of (2.12) is estimated in accordance with Theorem 1.

Let w be a sufficiently smooth function, specified in Q , and such that $w|_{\Gamma} = 0$. Since the function $\tilde{w}^h(t) - \hat{w}^h(t)$ is non-zero only in the triangles of which at least one vertex lies on S^h , we have

$$\begin{aligned} \|\nabla(\tilde{w}^h - \hat{w}^h)\|_{0,Q^h} \leq C\tau h^{-1} \|w\|_{2,1,Q} + \left\{ \sum_{(m) \in r^h} \int_0^T |w^h(x_m, y_m, t) - w(x_m, y_m, t)|^2 dt \right\}^{1/2} + \left\{ \sum_{(m) \in r^h} \int_0^T |w(x_m, y_m, t)|^2 dt \right\}^{1/2}. \end{aligned} \quad (2.14)$$

The following estimates for the second and third terms on the right-hand side of (2.14):

$$\left\{ \sum_{(m) \in r^h} \int_0^T |w^h(x_m, y_m, t) - w(x_m, y_m, t)|^2 dt \right\}^{1/2} \leq Ch \|w\|_{2,1,Q}, \quad (2.15)$$

$$\left\{ \sum_{(m) \in r^h} \int_0^T |w(x_m, y_m, t)|^2 dt \right\}^{1/2} \leq Ch \|w\|_{2,1,Q} \quad (2.16)$$

were in fact obtained in [8].

From (2.14)–(2.16) we have

$$\|\nabla(\tilde{w}^h - \hat{w}^h)\|_{0,Q^h} \leq C(h + \tau h^{-1}) \|w\|_{2,1,Q}. \quad (2.17)$$

We have proved this inequality for a sufficiently smooth function w . It obviously also holds for $u \in W_{2,1}^{2,1}(Q)$.

We can show in the same way that

$$\sup_{0 \leq t \leq T} \|\tilde{u}^h(t) - \hat{u}^h(t)\|_{0,Q^h} \leq C(h + \tau^{1/2}) \|u\|_{2,1,Q}. \quad (2.18)$$

Inequality (2.10) follows from (2.12), (2.13), (2.17), (2.18) and (2.1). The proof of inequality (2.11) is similar. The theorem is proved.

3. Convergence rate estimates

Let us now consider the rate of convergence of the approximate to the exact solution. Here we have:

Theorem 3

Assume that the conditions (1.5) and the estimate (1.6) hold; then, for $\tau = O(h^2)$, assuming that h is sufficiently small, we have

$$\|u - \tilde{v}\|_Q \leq C(T) h \|f\|_{0,Q}, \quad (3.1)$$

$$\|u - \tilde{v}\|_Q \leq C(T) h \|f\|_{0,Q}, \quad (3.2)$$

where $C(T)$ is a positive constant, dependent on T .

Proof. We start with the first initial-boundary value problem. Given any $\hat{\varphi} \in \dot{H}_{h, \tau}$ we have

$$\tau \sum_{n=1}^N \int_{\Omega} (u(t_n))_i \hat{\varphi}(t_n) d\Omega + \mathcal{L}_1(u, \hat{\varphi}) = \int_{\Omega} f \hat{\varphi} dQ. \quad (3.3)$$

Denote the function $\hat{u}^h - \hat{v}$ by \hat{w} . Taking \hat{w} as $\hat{\varphi}$, we obtain from (1.9) and (3.3):

$$\begin{aligned} & \tau \sum_{n=1}^N \int_{\Omega} (\hat{w}(t_n))_i \hat{w}(t_n) d\Omega + \mathcal{L}_1(\hat{w}, \hat{w}) \\ &= \tau \sum_{n=1}^N \int_{\Omega} (\hat{u}^h(t_n) - u(t_n))_i \hat{w}(t_n) d\Omega + \mathcal{L}_1(\hat{u}^h - u, \hat{w}). \end{aligned} \quad (3.4)$$

We transform the first term on the left-hand side of (3.4):

$$\begin{aligned} & \tau \sum_{n=1}^N \int_{\Omega} (\hat{w}(t_n))_i \hat{w}(t_n) d\Omega \\ &= \frac{1}{2} \int_{\Omega} |\hat{w}(t)|^2 d\Omega + \frac{\tau_2}{2} \sum_{n=1}^N \int_{\Omega} |(\hat{w}(t_n))_i|^2 d\Omega. \end{aligned} \quad (3.5)$$

We find an upper bound for the first term on the right-hand side of (3.4):

$$\begin{aligned} & \tau \sum_{n=1}^N \int_{\Omega} (\hat{u}^h(t_n) - u(t_n))_i \hat{w}(t_n) d\Omega = \int_{\Omega} (\hat{u}^h(T) - u(T)) \hat{w}(T) d\Omega \\ & - \tau \sum_{n=1}^N \int_{\Omega} (\hat{u}^h(t_{n-1}) - u(t_{n-1})) (\hat{w}(t_n))_i d\Omega \\ & \leq \varepsilon_1^{-1} |u - \hat{u}^h|_Q + \varepsilon_1 \int_{\Omega} |\hat{w}(T)|^2 d\Omega + \varepsilon_2^{-1} \sum_{n=1}^N \int_{\Omega} |u(t_n) - \hat{u}^h(t_n)|^2 d\Omega \\ & + \varepsilon_2 \tau^2 \sum_{n=1}^N \int_{\Omega} |(\hat{w}(t_n))_i|^2 d\Omega. \end{aligned} \quad (3.6)$$

A bound is easily obtained for the second term on the right-hand side of (3.4):

$$\mathcal{L}_1(\hat{u}^h - u, \hat{w}) \leq C \varepsilon^{-1} |u - \hat{u}^h|_Q^2 + \varepsilon |\hat{w}|_Q^2. \quad (3.7)$$

Putting $\varepsilon_1 = 1/4$ and $\varepsilon_2 = 1/2$ in (3.6), we find from (3.4)–(3.7) that

$$\begin{aligned} & \frac{1}{4} \int_{\Omega} |\hat{w}(T)|^2 d\Omega + \mathcal{L}_1(\hat{w}, \hat{w}) \leq C \sum_{n=1}^N \int_{\Omega} |u(t_n) - \hat{u}^h(t_n)|^2 d\Omega \\ & + C(1 + \varepsilon^{-1}) |u - \hat{u}^h|_Q^2 + \varepsilon |\hat{w}|_Q^2. \end{aligned} \quad (3.8)$$

The following bound can be obtained for the first time on the right-hand side of (3.8):

$$\begin{aligned} \sum_{n=1}^N \int_{\Omega} |u(t_n) - \hat{u}^h(t_n)|^2 d\Omega &= \tau^{-1} \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \int_{\Omega} |u(t_n) - \hat{u}^h(t)|^2 d\Omega dt \\ &\leq C\tau \|u\|_{2,1,Q} + C\tau^{-1} \|u - \hat{u}^h\|_{0,Q}. \end{aligned} \quad (3.9)$$

From (3.8) and (3.9) we obtain

$$\begin{aligned} \frac{1}{4} \int_{\Omega} |\hat{w}(T)|^2 d\Omega + \mathcal{L}_1(\hat{w}, \hat{w}) &\leq \varepsilon |\hat{w}|_Q^2 + C(\tau \|u\|_{2,1,Q}^2 \\ &+ \tau^{-1} \|u - \hat{u}^h\|_{0,Q} + (1 + \varepsilon^{-1}) |u - \hat{u}^h|_Q^2) = I. \end{aligned} \quad (3.10)$$

It can easily be shown that

$$\mathcal{L}_1(\hat{w}, \hat{w}) \geq C_1 \|\nabla \hat{w}\|_{0,Q}^2 - C_2 \|\hat{w}\|_{0,Q}^2. \quad (3.11)$$

From (3.10) and (3.11) we get

$$\int_{\Omega} |\hat{w}(T)|^2 d\Omega + \|\nabla \hat{w}\|_{0,Q}^2 \leq C_3 \|\hat{w}\|_{0,Q}^2 + C_4 I. \quad (3.12)$$

Using similar arguments, we can show that, for any t_n , $n = 1, 2, \dots, N$, we have

$$\int_{\Omega} |\hat{w}(t_n)|^2 d\Omega \leq C_3 \int_0^{t_n} \int_{\Omega} |\hat{w}(t)|^2 d\Omega dt + C_4 I. \quad (3.13)$$

Using (3.12) and (3.13), we can easily see that

$$|\hat{w}|_Q \leq C(T) I^{1/2}.$$

On making a suitable choice of ε in the expression for I and using Theorem 2 and the last inequality, we finally get

$$|u - \hat{v}|_Q \leq |u - \hat{u}^h|_Q + |\hat{w}|_Q \leq C(T) h \|u\|_{2,1,Q} \leq C(T) h \|f\|_{0,Q}.$$

The estimate (3.2) can be proved similarly. The theorem is proved.

Notes. 1. For the first initial-boundary value problem, the v.d.s. can be written in a somewhat different way: as the approximate solution we take a function $\hat{v} \in \hat{H}_{h,\tau}$, which satisfies the equation

$$\tau(v_{mn})_i \int_{\Omega} \varphi_m d\Omega + \mathcal{L}_1(\hat{v}, \varphi_{mn}) = \int_Q f \varphi_{mn} dQ \quad (3.14)$$

for all $(m) \in R^h$ and $n = 1, 2, \dots, N$. The convergence rate estimate is the same as before, but this latter scheme is more convenient.

2. Since the v.d.s's considered are implicit, the question of their numerical realization arises. Noting that $\tau = O(h^2)$, we can show that the systems of equations (1.9), (1.10) and (3.14) can be solved by the method of simple iterations with accuracy ϵ after $O(h^{-2} |\ln \epsilon h|)$ iterations. This subject will be dealt with in greater detail in later papers.

4. On the accuracy of the estimates

We shall consider approximate methods in which the approximate solution of the initial-boundary value problem, with right-hand side belonging to $L_2(Q)$, is sought as an element of some R -dimensional subspace in $V_2^{1,0}(Q)$, with a basis consisting of standard functions.

The set K of solutions of the initial boundary value problem with right-hand side belonging to the sphere $\|f\|_{0,Q} \leq 1$ is a compact set in $V_2^{1,0}(Q)$ [5].

Our problem lies in finding the accuracy to which the R -dimensional hyperplane approximates the compact set K , i.e., in obtaining Kolmogorov lower and upper bounds for the R -diameter $d_R(K)$ of K [10]:

$$d_R(K) = \inf_{L_R \subset V_2^{1,0}(Q)} \sup_{u \in K} \inf_{v \in L_R} \|u - v\|_Q,$$

where L_R is an R -dimensional subspace in $V_2^{1,0}(Q)$.

Since $\|u\|_{1,0,Q} \leq \|u\|_Q$ we have

$$d_R(K) \geq d_R^*(K), \quad (4.1)$$

where

$$d_R^*(K) = \inf_{L_R \subset W_2^{1,0}(Q)} \sup_{u \in K} \inf_{v \in L_R} \|u - v\|_{1,0,Q}.$$

We shall assume for simplicity that $T = \pi$ and that the domain Ω contains the rectangle $\{0 \leq x \leq \pi, 0 \leq y \leq \pi\}$. We denote by P the cube $\{0 \leq x \leq \pi, 0 \leq y \leq \pi, 0 \leq t \leq \pi\}$.

Notice that every subspace L_R defines in $W_2^{1,0}(P)$ a subspace, the dimensionality of which is not greater than the number R . Hence

$$d_R^*(K) \geq \inf_{L_R \subset W_2^{1,0}(P)} \sup_{u \in K} \inf_{v \in L_R} \|u - v\|_{1,0,P}. \quad (4.2)$$

We shall assume for simplicity that $R = MN$, where \sqrt{M} and N are integers.

We consider in $W_2^{1,0}(P)$ the subspace G , whose basis is formed by the functions $\sin kx \sin ly \sin nt$ for $k, l = \sqrt{M}, \sqrt{M}+1, \dots, 2\sqrt{M}$ and $n = 1, 2, \dots, N+1$.

It is easily shown by direct calculations that, for any $u \in G$

$$\|u\|_{1,0,P} \sim \left[\sum_{k,l=\sqrt{M}}^{2\sqrt{M}} \sum_{n=1}^{N+1} u_{kln}^2 (k^2 + l^2) \right]^{1/2},$$

$$\|u\|_{2,1,P} \sim \left\{ \sum_{k,l=\sqrt{M}}^{2/M} \sum_{n=1}^{N+1} u_{kln}^2 [(k^2+l^2)^2+n^2] \right\}^{1/2}, \quad (\text{cont'd})$$

where u_{klm} are the Fourier coefficients of the function u . Then,

$$\|u\|_{1,0,P} \geq C \left(\frac{M}{M^2+N^2} \right)^{1/2} \|u\|_{2,1,P}. \quad (4.3)$$

Consider an arbitrary subspace L_R in $W_2^{1,0}(P)$. Since the dimensionality of L_R is less than that of G , there will be an element u_{LR} in G which is orthogonal to L_R in $W_2^{1,0}(P)$.

Then, for any $v \in L_R$

$$\|u_{LR} - v\|_{1,0,P} \geq \|u_{LR}\|_{1,0,P}. \quad (4.4)$$

We continue the function u_{LR} into the whole of Q while preserving the class and norm of $W_2^{1,0}(P)$, in such a way that the function vanishes close to Γ and $u_{LR}|_{t=0}=0$. The function will then satisfy the initial and boundary conditions (1.2)–(1.4). We shall assume that $\|u_{LR}\|_{2,1,P} = (C_5 C_6)^{-1}$, where C_5 and C_6 are the constants in the inequalities

$$\|u\|_{2,1,Q} \leq C_5 \|u\|_{2,1,P}, \quad \|Lu\|_{0,Q} \leq C_6 \|u\|_{2,1,Q}.$$

It follows from what has been said that $u_{LR} \in K$.

Then we find from (4.3) and (4.4) that, given any subspace L_R in $W_2^{1,0}(P)$

$$\sup_{u \in K} \inf_{v \in L_R} \|u - v\|_{1,0,P} \geq \|u_{LR}\|_{1,0,P} \geq C \left(\frac{M}{M^2+N^2} \right)^{1/2}. \quad (4.5)$$

From (4.1), (4.2), and (4.5) we obtain the inequality

$$d_R(K) \geq C \left(\frac{M}{M^2+N^2} \right)^{1/2},$$

whence it follows that, for $M \sim N$,

$$d_R(K) \geq CR^{-1/4}.$$

An upper bound for the diameter is provided by the estimates (3.1) and (3.2), which likewise are of order $R^{-1/4}$.

In short, upper and lower bounds of the same order of accuracy have been obtained. Hence the estimates (3.1) and (3.2) are not improvable with respect to order.

Translated by D. E. Brown

REFERENCES

1. DOUGLAS, J., and DUPONT, T., Galerkin methods for parabolic equations, *SIAM J. Numer. Anal.*, 7, No. 4, 575-626, 1970.
2. SWARTZ, B., and WENDROFF, B., Generalized finite difference schemes, *Math. comput.*, 23, No. 105, 37-49, 1969.
3. ZLAMAL, M., Finite element methods for parabolic equations, *Math. Comput.*, 28, No. 126, 393-404, 1974.
4. ASTRAKHANTSEV, G. P., A finite difference method for solving the third boundary value problem for elliptic and parabolic equations in an arbitrary domain. Iterative solution of difference equations, 2., *Zh. vychisl. Mat. mat. Fiz.*, 11, No. 3, 677-687, 1971.
5. SLOBODETSKII, L. N., Generalized Sobolev spaces and their applications to boundary value problems for partial differential equations, *Uch. zap. Leningr. gos. ped. in-ta im. A. I. Gertsena*, 197, 54-112, 1958.
6. LADYZHENSKAYA, O. A., SOLONNIKOV, V. A., and URAL'TSEVA, N. N., *Linear and quasi-linear equations of parabolic type*, (Lineinye i kvazilineinye uravneniya parabolicheskogo tipa), Nauka, Moscow, 1967.
7. OGANESYAN, L. A., Convergence of difference schemes with improved approximation of the boundary, *Zh. vychisl. Mat. mat. Fiz.*, 6, No. 6, 1029-1042, 1966.
8. OBANESYAN, L. A., RIVKIND, V. YA., and RUKHOVETS, L. A., Variational difference methods for solving elliptic equations, Part I, in: *Differential equations and their applications* (Differents. ur-niya i ikh primeneniye), No. 5, Vilnius, 1973.
9. IL'IN, V. P., Some inequalities in functional spaces and their use in the study of the convergence of variational processes, *Tr. Matem. in-ta Akad. Nauk SSSR*, 53, 64-127, 1959.
10. KOLMOGOROFF, A., Über die beste Annäherung von Functionen einer gegebenen Functionenklasse, *Math. Ann.*, 37, 107-111, 1936.

ON NUMERICAL ISOLATION OF THE BOUNDED SOLUTIONS OF SYSTEMS OF LINEAR PARTIAL DIFFERENTIAL EQUATIONS OF THE EVOLUTIONARY TYPE*

SH. M. NASIBOV

Baku

(Received 24 February 1976)

NUMERICAL isolation of the bounded solutions is discussed for certain systems of linear partial differential equations of the evolutionary type. Boundedness of the solutions at infinity, or at singular points, where the coefficients become infinite, is taken as the boundary condition at the relevant points. The manifold of bounded solutions is isolated without investigating the complicated asymptotic behaviour of the particular solutions at these points. An applied problem is considered as an example.

1. Isolation of the solutions, bounded at infinity

1. We consider a linear differential operator acting on vector functions u of two independent variables x and t :

$$P_{x,t}[u] = H_2 \left[\frac{\partial}{\partial t} \right] u - \mathcal{L}_2 \left[\frac{\partial}{\partial x} \right] u, \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_q \end{pmatrix},$$

**Zh. vychisl. Mat. mat. Fiz.*, 17, 1, 119-135, 1977.

where $H_2[\partial/\partial t]$ is a linear differential operator of the form $H_2[\partial/\partial t] = A_0 \partial^2 / \partial t^2 + A_1 \partial / \partial t$, A_0, A_1 are constant square matrices of order q , and $\mathcal{L}_2[\partial/\partial x]$ is a second-order linear differential operator of the form $\mathcal{L}_2[\partial/\partial x] = \partial^2 / \partial x^2 + B(x)$, where $B(x)$ is a square matrix of order q , whose elements depend only on the one independent variable x .

Suppose that we want to find the vector function u , representing a solution of the equation

$$P_{x,t}[u] = -f(x)\mu(t) \quad (1.1)$$

in $\Omega_{x,t} = R_x\{x | a < x < \infty\} \times (0, T)$, where T is arbitrary, under the initial conditions

$$u(x, 0) = \varphi_0(x), \quad (\partial u / \partial t) |_{t=0} = \varphi_1(x). \quad (1.2)$$

At the left-hand end $x = a$ of the semi-infinite interval, the function $u(x, t)$ satisfies a linear boundary condition, and as $x \rightarrow \infty$, the condition

$$\sup_{t \geq 0} |\exp(-\gamma_k t) u_k(x, t)| = O(1), \quad (1.3)$$

where γ_k are non-negative constants. We assume that the matrix $B(x)$, the initial conditions, and the function $f(x)$, are continuous functions in $[a, \infty)$.

When integrating problem (1.1)–(1.3) numerically, the question arises of the correct translation of condition (1.3) from infinity to a finite point.

We shall assume that $\mu(t)$, u , $\partial u / \partial x$, $\partial^2 u / \partial x^2$, regarded as functions of t , are originals. We denote the image of u by

$$v(x, p) = \int_0^\infty e^{-pt} u(x, t) dt,$$

and the image of $\mu(t)$ by

$$M(p) = \int_0^\infty e^{-pt} \mu(t) dt.$$

We transform to images in (1.1). We obtain as a result the systems of ordinary differential equations

$$d^2 v(x, p) / dx^2 + [B(x) - H_2(p)] v = -g(x, p), \quad (1.4)$$

where $g(x, p) = F(x, p) + F_0(x, p)$, $F(x, p) = f(x)M(p)$, $F_0 = A_0 \varphi_0 p + A_0 \varphi_1 + A_1 \varphi_0$, $H_2(p) = A_0 p^2 + A_1 p$, with the condition

$$|v(x, p)| = O(1) \text{ as } x \rightarrow \infty \text{ uniformly with respect to } p. \quad (1.5)$$

The matrix $B(x)$, the functions $\varphi_0(x)$, $\varphi_1(x)$ and $f(x)$ have specified asymptotic forms as $x \rightarrow \infty$:

$$\begin{aligned}
 B(x) &\sim \sum_{h=0}^{\infty} \frac{B_h}{x^h}, \\
 \varphi_j(x) &\sim \sum_{h=0}^{\infty} \frac{\varphi_{jh}}{x^h}, \quad j=0, 1, \\
 f(x) &\sim \sum_{h=0}^{\infty} \frac{f_h}{x^h}.
 \end{aligned} \tag{1.6}$$

It follows from [1] that, for large x , the manifolds of bounded solutions of system (1.4)–(1.6) are described by the equation

$$dv(x, p) / dx = \alpha(x, p)v + \beta(x, p). \tag{1.7}$$

Here, for α we have

$$d\alpha / dx + \alpha^2 + B(x) - H_2(p) = 0, \tag{1.8}$$

$$\alpha|_{x \rightarrow \infty} \rightarrow \alpha_0, \tag{1.9}$$

$$\alpha(x, p) \sim \sum_{h=0}^{\infty} \frac{\alpha_h(p)}{x^h} \quad \text{as } x \rightarrow \infty. \tag{1.10}$$

The α_0 is a square matrix, dependent on the complex parameter p , such that $\alpha_0^2 = H_2(p) - B_0$; all its eigenvalues must have negative real parts. We shall see later that such a choice is possible and is unique. Further, we have for $\beta(x, p)$:

$$d\beta / dx + \alpha(x, p)\beta(x, p) = g(x, p), \tag{1.11}$$

$$\beta|_{x \rightarrow \infty} \rightarrow \beta_0 = \alpha_0^{-1}g_0,$$

$$\beta(x, p) \sim \sum_{h=0}^{\infty} \frac{\beta_h(p)}{x^h} \quad \text{as } x \rightarrow \infty, \tag{1.12}$$

where $g_0 = M(p)f_0 + A_0\varphi_0, 0p + A_0\varphi_1, 0 + A_1\varphi_0, 0$.

The matrices α_k and β_k are determined formally from the recurrence relations

$$\sum_{l+j=m; l, j \geq 0} \alpha_l \alpha_j + B_m = (m-1)\alpha_{m-1}, \tag{1.13}$$

$$\alpha_0 \beta_m = g_m + (m-1)\beta_{m-1} - \sum_{l+j=m; l, j \geq 0} \alpha_l \beta_j - \beta_0 \alpha_m, \tag{1.14}$$

$$g_m = M(p)f_m + A_0\varphi_0, mp + A_0\varphi_1, m + A_1\varphi_0, m, \quad m=1, 2, \dots$$

It follows immediately from (1.7) that the boundary value problem in (a, ∞) for system (1.4) with the given boundary condition at the point $x = a$ and the condition $|v(x, p)| \rightarrow 0$ as $x \rightarrow \infty$ can be reduced to the equivalent problem in $[a, x_\infty]$, where, as the boundary condition at the point x_∞ , we write the equation of the manifold (1.7), stable as $x \rightarrow \infty$. After Laplace inversion, the relation obtained in (1.7), if it exists, describes the nature of the manifold of stable solutions as $x \rightarrow \infty$ of Eq. (1.1), which has an irregular singularity at infinity. Because of this, condition (1.3) for Eq. (1.1) can be effectively "displaced" from infinity to a finite point x_∞ and the above boundary value problem (1.1)–(1.3) in $[a, \infty)$ reduces to the equivalent problem in $[a, x_\infty]$.

2. We will consider the case when, in Eq. (1.1), $A_0 = 0$, $A_1 = E$, where E is the unit matrix. With a view to reducing the amount of calculations, we put $f = 0$, $\phi_0 = 0$. (The case $f \neq 0$, $\phi_0 \neq 0$ is treated in a similar way). In short, we consider the system of linear parabolic equations

$$\frac{\partial u_j}{\partial t} = \frac{\partial^2 u_j}{\partial x^2} + \sum_{h=1}^q B_{jh}(x) u_h, \quad 1 \leq j \leq q. \quad (1.15)$$

We assume that all the eigenvalues $\{\lambda_1, \dots, \lambda_q\}$ of the matrix $B_0 = B(\infty)$ are simple. We can then assume without loss of generality that B_0 is a diagonal matrix, with $\{\lambda_1, \dots, \lambda_q\}$ along the diagonal.

In the case of the system obtained by Laplace inversion from (1.15), Eq. (1.8) and condition (1.9) become

$$\alpha' + \alpha^2 + B(x) - pE = 0, \quad \alpha|_{x \rightarrow \infty} \rightarrow \alpha_0. \quad (1.16)$$

Here, α_0 is found from the matrix equation $\alpha_0^2 = pE - B_0$, containing the complex parameter p , with a condition such that the eigenvalues have a negative real part. Denote by ρ_0 the maximum real part of any eigenvalue of the matrix B_0 , i.e., we put

$$\rho_0 = \max_{1 \leq h \leq q} \operatorname{Re} \lambda_h.$$

As functions of the complex argument p , the matrix elements $[\alpha_0(p)]_{ij} = -(p - \lambda_i)^{-1/2} \delta_{ij}$, $1 \leq i, j \leq q$, are defined for all values of p in the right-hand half-plane $\operatorname{Re} p > \rho_0$, they have no branching points, and they have a negative real part (below, $W^{1/2}$, $\operatorname{Re} W > 0$, is always the value of the root which lies in the right-hand half-plane). After this unique choice of $\alpha_0(p)$, it is easily shown, by arguments similar to those employed in [1], that Eq. (1.16) with condition (1.14) has a unique solution for all values of the complex parameter in the right-hand half-plane

$$\operatorname{Re} p > \omega_0, \quad \omega_0 = \max\{\rho_0, \max_{1 \leq h \leq q} \gamma_h\},$$

and for large x , this solution can be expanded in the asymptotic series (1.10), whose coefficients are formally defined from (1.13). On determining $\alpha_m(p)$ successively from the recurrence relations (1.13), we can show by induction that

$$\alpha_m(p) = \sum_{h=1}^m C_h \alpha_0^{-h}(p),$$

where C_k are constant matrices, and $m = 1, 2, \dots$. The equation of the manifold of solutions, bounded as $x \rightarrow \infty$, of the system obtained by Laplace inversion of (1.15), takes the form, after multiplying on the left by the inverse matrix $\alpha_0^{-1}(p)$:

$$\alpha_0^{-1}(p) \frac{dv(x, p)}{dx} = -v(x, p) + \left(\sum_{m=1}^{\infty} \frac{\alpha_m(p)}{x^m} \right) v(x, p), \quad (1.17)$$

where

$$\alpha_m(p) = \sum_{k=2}^{m+1} c_k \alpha_0^{-k}(p).$$

Since

$$\det \|\alpha_0(p)\| = \prod_{1 \leq k \leq q} (p - \lambda_k)^{1/k} \neq 0$$

for all p in the half-plane $\operatorname{Re} p > \omega_0$, then $\alpha_0^{-1}(p)$ exists. We can show directly that the matrices $\alpha_0^{-k}(p)$, of the form

$$[\alpha_0^{-k}(p)]_{jl} = (p - \lambda_j)^{-k/2} \delta_{jl}, \quad 1 \leq j, l \leq q,$$

possess the matrix-originals $\mathcal{R}_k(t)$ for $\operatorname{Re} p > \rho_0$, where

$$[\mathcal{R}_k(t)]_{jl} = \frac{\exp(\lambda_j t) (t - \lambda_j)^{(k-2)/2}}{\Gamma(k/2)} \delta_{jl}, \quad 1 \leq j, l \leq q.$$

Noting that $\partial v(x, p)/\partial x$ and $v(x, p)$ are images, with

$$\operatorname{Re} p > \max_{1 \leq k \leq q} \gamma_k,$$

of the vector-function-originals $\partial u/\partial x$ and $u(x, t)$, respectively, and using the theorem on the multiplication of images, which holds for $\operatorname{Re} p > \omega_0 = \max\{\rho_0, \max \gamma_k\}$, we obtain from (1.17):

$$\mathcal{R}_0(t) * \frac{\partial u}{\partial x} = -u + \mathcal{R}(x, t) * u(x, t),$$

where

$$\mathcal{R}(x, t) \sim \sum_{m=1}^{\infty} \frac{\mathcal{R}_m(t)}{x^m},$$

and the symbol $*$ denotes the convolution of the functions $g(t)$ and $h(t)$:

$$g(t) * h(t) = \int_0^t g(t-\tau) h(\tau) d\tau.$$

We now take the case when the eigenvalues of the matrix B_0 are not simple. Let their multiplicities be ν_1, \dots, ν_k ; $\nu_1 + \dots + \nu_k = q$. Using the same arguments as in the case when all the eigenvalues are simple, we can assume that B_0 has the Jordan form

$$B_0 = \sum_{\nu_1 + \dots + \nu_k = q} \oplus J_{\nu_l},$$

where $J_{\nu_1}, \dots, J_{\nu_k}$ are the lower triangular Jordan cells, corresponding to the multiplicities ν_1, \dots, ν_k of the eigenvalues $\lambda_1, \dots, \lambda_k$. The matrix $\alpha_0(p)$ is a single-valued analytic function of

the complex variable p in the half-plane

$$\operatorname{Re} p > \rho_0 = \max_{1 \leq h \leq q} \lambda_h$$

and has the form

$$\alpha_0(p) = - \sum_{v_1 + \dots + v_k = q} \oplus I_{v_l},$$

where $I_\rho, \rho = v_1, \dots, v_k$, is a triangular matrix of the form

$$\begin{vmatrix} (p - \lambda_\rho)^{1/2} & & & 0 \\ 2^{-1}(p - \lambda_\rho)^{-1/2} & (p - \lambda_\rho)^{1/2} & & \\ \dots & \dots & \dots & \dots \\ 2^{-\rho}(2\rho - 3)!(p - \lambda_\rho)^{(2\rho-1)/2} & \dots & 2^{-1}(p - \lambda_\rho)^{-1/2} & (p - \lambda_\rho)^{1/2} \end{vmatrix},$$

and it has an inverse $\alpha_0^{-1}(p) = I_{v_1}^{-1} \oplus \dots \oplus I_{v_k}^{-1}$, since

$$\det \|\alpha_0(p)\| = \prod_{v_1 + \dots + v_k = q} (p - \lambda_{v_l})^{1/2}$$

nowhere vanishes in the half-plane $\operatorname{Re} p > \rho_0$. It is easily shown that

$$[I_\rho^{-1}]_{ij} = \frac{[I_\rho]_{ij}}{\det \|I_\rho\|} = C_{ij} \left(\frac{1}{p - \lambda_\rho} \right)^{(2l-2j+1)/2},$$

where $1 \leq l \leq \rho, 1 \leq j \leq l, \rho = v_1, \dots, v_k, C_{ij}$ are certain constants.

In the same way, we have

$$\alpha_0^{-m}(p) = \sum_{v_1 + \dots + v_k = q} \oplus I_{v_l}^{-m},$$

where

$$[I_\rho^{-m}]_{ij} = C_{ij} \left(\frac{1}{p - \lambda_\rho} \right)^{(2l-2j+1)/2}, \quad 1 \leq l \leq \rho, \quad 1 \leq j \leq l, \quad \rho = v_1, \dots, v_k,$$

since $\det \|\alpha_0^{-1}(p)\| \neq 0$ in the half-plane $\operatorname{Re} p > \rho_0$. Obviously, for $\operatorname{Re} p > \rho_0$, the matrices $\alpha_0^{-k}(p), k=1, 2, \dots$, have the matrix-originals

$$\alpha_0^{-k}(p) \doteq \mathcal{R}_k(t) = \mathcal{R}_{k, v_1} \oplus \dots \oplus \mathcal{R}_{k, v_k},$$

where

$$\begin{aligned} \mathcal{R}_{k\rho}(t) \|_{ij} &= \text{const} \exp(\lambda_\rho t) (t - \lambda_\rho)^{(2l-2j+k-2)/2}, \quad 1 \leq l \leq \rho, \\ 1 \leq j \leq l, \quad \rho &= v_1, \dots, v_k. \end{aligned}$$

Hence, we can see that, in the half-plane

$$\operatorname{Re} p > \omega_0 = \max \{\gamma, d\},$$

where

$$d = \max_{1 \leq l \leq q} \left[\operatorname{Re} \|B(\infty)\|_{ll} + \sum_{j=1, l \neq j}^q |B^{lj}(\infty)| \right] \geq \rho_0 \geq \max_{1 \leq h \leq q} \operatorname{Re} \lambda_h,$$

and γ is the growth exponent with respect to t of the function $u(x, t)$, representing the solution, bounded as $x \rightarrow \infty$, of the system (1.15) with the appropriate initial and boundary data, it is possible to pass from (1.17) to originals.

3. Let us return to Sec. 1. We assume for simplicity that the matrices A_0, A_1 , and B_0 have the diagonal form. Obviously, if $\operatorname{Re} A_0 > 0$, then a positive constant ω_1 exists, such that, in the domain $\operatorname{Re} p > \omega_1$, the two-valued expression $[H_2(p) - B_0]^{1/2}$ can be divided, on the one hand into a regular branch with positive real part $\operatorname{Re} \{[H_2(p) - B_0]^{1/2}\} > 0$, and on the other hand, into elements of the matrix $D_l(p) = \|H_2(p) - B_0\|_l = H_2^{(l)}(p) - B_0^{(l)}, 1 \leq l \leq q$, which do not vanish. Hence the matrices $\alpha_0^{-k}(p)$ exist for $\operatorname{Re} p > \omega_1$, since $\det \|\alpha_0(p)\| = \prod D_l(p) \neq 0$, and they have the form $\|\alpha_0^{-k}(p)\|_l = (-1)^k [D_l(p)]^{-1}, 1 \leq l \leq q$. In addition, with $\operatorname{Re} p > \omega_1$, the elements of the matrices $\alpha_0^{-k}(p), k=1, 2, \dots$, are analytic functions with no zeros, they tend to zero as $|p| \rightarrow \infty$ uniformly with respect to $\arg p$, and the integral

$$\int_{\omega - i\infty}^{\omega + i\infty} |\alpha_0^{-k}(p)| dp$$

exists for all $k = 1, 2, \dots$; hence $\alpha_0^{-k}(p)$ are the images of matrices $\mathcal{R}_k(t)$, where

$$\|\alpha_0^{-k}(p)\|_l = \|\mathcal{R}_k(t)\|_l = (-1)^k \int_{\omega - i\infty}^{\omega + i\infty} \frac{e^{pt} dp}{[D_l(p)]^{k/2}}, \quad 1 \leq l \leq q.$$

Finally, it must be mentioned that, after multiplying by $\alpha_0^{-1}(p)$, we can pass to images in (1.7) in the half-plane $\operatorname{Re} p > \omega_0 = \max \{\omega_1, \gamma\}$, provided that we can justify the operation

$$\mathcal{L}^{-1} \left[\left(\sum_{m=1}^{\infty} \frac{\alpha_m(p)}{x^m} \right) v(x, p) \right] = \sum_{m=1}^{\infty} \frac{1}{x^m} \mathcal{L}^{-1} [\alpha_m(p) v(x, p)], \quad (1.18)$$

where \mathcal{L}^{-1} is the inverse Laplace transform. To this end, we shall prove:

Lemma

Given any fixed $x \in [a, \infty)$ let the functions $F(x, p), v(x, p)$, and every term of the functional sequence $\{c_k(p)\}$, regarded as functions of the complex argument p , be uniquely defined in the half-plane $M_p = \{p | \operatorname{Re} p > \omega_0\}$. Further, suppose that:

a) the series $\sum_{k=0}^{\infty} \frac{c_k(p)}{x^k}$ is the asymptotic (as $x \rightarrow \infty$) series of the function $F(x, p)$, specified in $\Omega_{xp} = [a, \infty) \times M_p$ uniformly with respect to p ;

b) the integrals

$$\int_{\omega - i\infty}^{\omega + i\infty} v(x, p) e^{pt} dp, \quad \int_{\omega - i\infty}^{\omega + i\infty} W(x, p) e^{pt} dp,$$

where $W(x, p) = F(x, p) v(x, p)$, and

$$\int_{\omega - i\infty}^{\omega + i\infty} c_k(p) e^{pt} dp,$$

where $k = 1, 2, \dots$, exist for $\operatorname{Re} p > \omega_0$;

c) the function

$$u(x, t) = (2\pi i)^{-1} \int_{\omega - i\infty}^{\omega + i\infty} v(x, p) e^{pt} dp$$

is bounded for all $x \in [a, \infty)$ for any fixed $t \in (0, T)$, where T is an arbitrary number.

We then have

$$\int_{\omega - i\infty}^{\omega + i\infty} W(x, p) e^{pt} dp = \sum_{h=0}^{\infty} x^{-h} \int_{\omega - i\infty}^{\omega + i\infty} c_h(p) v(x, p) e^{pt} dp \quad (1.19)$$

as $x \rightarrow \infty$ uniformly with respect to $t \in (0, T)$, where T is arbitrary.

Proof. We have to show that the relation

$$\int_{\omega - i\infty}^{\omega + i\infty} W(x, p) e^{pt} dp = \sum_{h=0}^N x^{-h} \int_{\omega - i\infty}^{\omega + i\infty} c_h(p) v(x, p) e^{pt} dp + o(x^{-N})$$

holds for any integer $N \geq 0$ and for all $t \in (0, T)$, where T is arbitrary.

From condition a) we have

$$F(x, p) = \sum_{h=1}^N \frac{c_h(p)}{x^h} + o(x^{-N}),$$

which holds for any integer $N \geq 0$, and indeed holds uniformly with respect to $p \in M_p$.

Multiplying this relation by $v(x, p) e^{pt}$, $t > 0$, and integrating along the imaginary axis $\operatorname{Im} p = \theta$, lying to the right of $\operatorname{Re} p = \omega_0$, we obtain

$$\int_{\omega - i\infty}^{\omega + i\infty} W(x, p) e^{pt} dp = \sum_{h=1}^N \frac{1}{x^h} \int_{\omega - i\infty}^{\omega + i\infty} c_h(p) v(x, p) e^{pt} dp + o(x^{-N}) u(x, t). \quad (1.20)$$

The required relation (1.19) follows from conditions b) and c) and relation (1.20).

Application of the lemma gives us (1.18). Hence we have:

Theorem

Assume that

a) $B(x)$, $\varphi_0(x)$, $\varphi_1(x)$ and $f(x)$ in (1.1)–(1.3) have the given asymptotic forms (1.6) as $x \rightarrow \infty$;

b) $\operatorname{Re} A_0 > 0$, $\mu_k(t) = O(e^{\gamma_{3k}t})$ as $t \rightarrow \infty$ (γ_{3k} are positive constants), and there exists

$$\mathcal{L}_t[\mu(t)] = \int_0^\infty e^{pt} \mu(t) dt;$$

c) the solution of problem (1.1)–(1.3) $u(x, t)$ is such that, as $t \rightarrow \infty$, it has a bounded degree of growth

$$\sup_{x \geq a} \left| \frac{\partial^l u_k}{\partial x^l} \right| = O(\exp(\gamma_{lk} t))$$

and there exists $\mathcal{L}_t[\partial^l u_k / \partial x^l]$, where $l=0, 1, 2, 1 \leq k \leq q$, γ_{lk} are positive constants.

Then, for any $t \in (0, T)$, T is arbitrary, the equation of the manifold of solutions, bounded as $x \rightarrow \infty$, of problem (1.1)–(1.3) is

$$\mathcal{R}_0(t) * \frac{\partial u}{\partial x} = -u(x, t) + \mathcal{R}(x, t) * u(x, t) + \Gamma(x, t),$$

where $\mathcal{R}(x, t)$ and $\Gamma(x, t)$ have the asymptotic expansions

$$\mathcal{R}(x, t) \sim \sum_{m=1}^{\infty} \frac{\mathcal{R}_m(t)}{x^m}, \quad \Gamma(x, t) \sim \sum_{m=0}^{\infty} \frac{\Gamma_m(t)}{x^m},$$

and $\mathcal{R}_0(t)$, $\mathcal{R}_m(t)$ and $\Gamma_m(t)$ are defined, for

$$\operatorname{Re} p > \omega_0 = \max \{ \omega_1, \max_{0 \leq l \leq 3} \max_{1 \leq k \leq q} \gamma_{lk} \}$$

by the integrals

$$\begin{aligned} \|\mathcal{R}_0(t)\|_{ij} &= (2\pi i)^{-1} \int_{\omega-i\infty}^{\omega+i\infty} \|\alpha_0^{-1}(p)\|_{ij} e^{pt} dp, \\ \|\mathcal{R}_m(t)\|_{ij} &= (2\pi i)^{-1} \int_{\omega-i\infty}^{\omega+i\infty} \|\alpha_0^{-1}(p) \alpha_m(p)\|_{ij} e^{pt} dp, \\ \|\Gamma_m(t)\|_l &= (2\pi i)^{-1} \int_{\omega-i\infty}^{\omega+i\infty} \|\alpha_0^{-1}(p) \beta_m(p)\|_l e^{pt} dp, \quad 1 \leq l, j \leq q. \end{aligned} \quad (1.21)$$

The matrices $\alpha_m(p)$ and $\beta_m(p)$ are formally defined by the recurrence relations (1.13), (1.14). The series

$$\sum_{m=1}^{\infty} \frac{\mathcal{R}_m(t)}{x^m} \text{ and } \sum_{m=0}^{\infty} \frac{\Gamma_m(t)}{x^m}$$

are asymptotically convergent as $x \rightarrow \infty$, for any fixed $t \in (0, T)$, where T is arbitrary.

Proof. The Laplace transform of $Y_0(x, t) = \mathcal{R}_0(t) * \partial u / \partial x + u - \mathcal{R} * u - \Gamma$, where $u(x, t)$ is a solution of (1.1)–(1.3), exists for $\operatorname{Re} p > \omega_0$; performing the transformation, we get

$$W_0(x, p) = \alpha_0^{-1} \frac{\partial v}{\partial x} + v - \eta v - \xi, \quad W_0(x, p) \doteq Y_0(x, p),$$

where

$$\eta(x, p) = \sum_{m=1}^{\infty} \frac{\alpha_0^{-1}(p) \alpha_m(p)}{x^m}, \quad \xi(x, p) = \sum_{m=0}^{\infty} \frac{\alpha_0^{-1}(p) \beta_m(p)}{x^m}.$$

The expressions

$$W(x, p) = \alpha_0 W_0 = \frac{\partial v}{\partial x} - \alpha v - \beta, \quad \alpha = \sum_{m=0}^{\infty} \frac{\alpha_m}{x^m}, \quad \beta = \sum_{m=0}^{\infty} \frac{\beta_m}{x^m},$$

represent the asymptotic solutions of (1.7), (1.8) and (1.11), (1.12) as $x \rightarrow \infty$; for them, we obtain from (1.7) the equation $\partial W / \partial x + \alpha W = 0$, containing the complex parameter p , $\operatorname{Re} p > \omega_0$.

Let $u(x, t)$ be a solution, bounded as $x \rightarrow \infty$, of the problem (1.1)–(1.3); then $v(x, p)$ is also bounded as $x \rightarrow \infty$, for any p , $\operatorname{Re} p > \omega_0$. Then, using Lemma 1.1 of [1], for all p , $\operatorname{Re} p > \omega_0$, we have $W(x, p) = \alpha_0 W_0 = 0$. Since $\alpha_0(p)$ never vanishes in the right-hand half-plane $\operatorname{Re} p > \omega_0$, we have $W_0(x, p) = 0$, and hence $Y_0(x, t) \equiv 0$ as $x \rightarrow \infty$, for any t . Hence any solution, bounded as $x \rightarrow \infty$, of the problem (1.1)–(1.3) appears in (1.21). Conversely, if $W(x, p) \equiv 0$ for any $x \in [a, \infty)$ for all p of the half-plane $\operatorname{Re} p > \omega_0$, then $v(x, p)$ will satisfy the equation $\partial v / \partial x = \alpha v + \beta$, all the solutions of which, by Theorem 3.1 of [2], Chapter 13, and Theorem 2 of [3], Chapter 2, are bounded as $x \rightarrow \infty$ for any value of the complex parameter p , $\operatorname{Re} p > \omega_0$. It follows from this that all the solutions (1.21) are bounded as $x \rightarrow \infty$, and hence they are solutions, bounded as $x \rightarrow \infty$, of the problem (1.1)–(1.3) for all $t \in (0, T)$, where T is arbitrary. The theorem is proved.

2. Isolation of the solutions, bounded in the neighbourhood of a singular point

1. Suppose that we wish to find the solution of Eq. (1.1), bounded in the interval $[0, d]$, with initial data (1.2) and a linear boundary condition at $x = d$, when $B(x)$ has a given asymptotic expansion as $x \rightarrow 0$, e.g., of the form

$$B(x) \sim \frac{B_{-2}}{x^2} + \frac{B_{-1}}{x} + \sum_{k=0}^{\infty} B_k x^k. \quad (2.1)$$

It follows from the results obtained in [4] that the condition for the solutions of the systems of ordinary differential equations (1.7) and (1.15), which have a regular singularity at the point $x = 0$, to be bounded for sufficiently small x is equivalent to the condition

$$x \partial v / \partial x = \alpha v + \beta(x, p). \quad (2.2)$$

Here, we have for α and β respectively:

$$\frac{\alpha^4}{x} + \frac{\alpha^2}{x^2} - \frac{\alpha}{x^2} + B(x) - H_2(p) = 0, \quad (2.3)$$

$$\alpha|_{x \rightarrow 0} \rightarrow \alpha_0, \quad (2.4)$$

$$\alpha(x, p) \sim \sum_{h=0}^{\infty} \alpha_h x^h \quad \text{as } x \rightarrow 0, \quad (2.5)$$

where α_0 is the root of the quadratic equation $\alpha_0^2 + \alpha_0 + B_{-2} = 0$.

We shall choose the root for which the eigenvalues have a positive real part. Such a choice always exists, and is unique, provided that $B_{-2} \leqslant 1/E$, where E is the unit matrix. Further,

$$\frac{\beta^4}{x} + \frac{\alpha\beta}{x^2} - \frac{\beta}{x^2} = g(x, p), \quad (2.6)$$

$$\beta|_{x \rightarrow 0} \rightarrow \beta_0 = (\alpha_0 - E)^{-1} g_{-2}, \quad (2.7)$$

$$\beta(x, p) \sim \sum_{h=0}^{\infty} \beta_h x^h \quad \text{as } x \rightarrow 0. \quad (2.8)$$

Here, $g_{-2} = M(p)f_{-2} + A_0\varphi_{0,-2}p + A_0\varphi_{1,-2} + A_1\varphi_{0,-2}$, where f_{-2} and $\varphi_{j,-2}$, $j=0, 1$, are the coefficients of x^{-2} in the expansion of the functions $\varphi_j(x)$, $j=0, 1$, and f , in the neighbourhood of zero, i.e., the following expansions are assumed to hold:

$$\varphi_j(x) = \frac{\varphi_{j,-1}}{x} + \frac{\varphi_{j,-2}}{x^2} + \sum_{h=0}^{\infty} \varphi_{j,h} x^h, \quad j=0, 1, \quad (2.9)$$

$$f(x) = \frac{f_{-2}}{x^2} + \frac{f_{-1}}{x} + \sum_{h=0}^{\infty} f_h x^h.$$

Employing similar arguments to those in [4], we see that (2.3), (2.4), and (2.6), (2.7) have a unique solution for all values of the complex parameter p in the half-plane $\operatorname{Re} p > \omega_0$, where the solution can be expanded in powers of x when x is small. The matrices α and β are formally found from (2.3)–(2.5) and (2.6)–(2.8), while the series (2.5), (2.8) are asymptotically convergent for small $x \rightarrow 0$ and any p , $\operatorname{Re} p > \omega_0$. After finding α and β , it remains to perform an inverse Laplace transformation in (2.2). The resulting relation effectively replaces for sufficiently small x the condition that the solutions of Eqs. (1.1), (1.2), (2.1), (2.9), having a regular singularity at the point $x = 0$, be bounded at this point.

2. Our method can be used to solve the boundary value problem

$$\frac{\partial u}{\partial t} = Lu, \quad Lu = \frac{\partial^2 u}{\partial x^2} + \frac{1}{x} \frac{\partial u}{\partial x}, \quad 0 < x < 1, \quad t > 0, \quad (2.10)$$

$$u|_{t=0}=0, \quad (2.11)$$

$$\partial u / \partial t + \gamma u|_{x=1} = \mu(t), \quad (2.12)$$

$$u(x, t) = O(1) \text{ as } x \rightarrow 0 \text{ for all } t > 0. \quad (2.13)$$

For Eqs. (2.10) and (2.14), relations (2.3), (2.4) take the form

$$\frac{d\alpha}{dx} + \frac{\alpha^2}{x^2} - p = 0, \quad (2.14)$$

$$\alpha|_{x \rightarrow 0} \rightarrow 0 \text{ for any } p, \operatorname{Re} p > \omega_0, \quad (2.15)$$

where ω_0 is the growth exponent with respect to t of the solution $u(x, t)$ of problem (2.10)–(2.13). From (2.14) and (2.15) we find the following recurrence relations for finding the coefficients in the asymptotic expansion:

$$\alpha_{2m}(p) = - \sum_{k=1}^{[m/2]} \alpha_{2k} \alpha_{2m-2k} - \frac{\alpha_m^2}{2m} \varepsilon_m, \quad \alpha_2(p) = \frac{p}{2}, \quad (2.16)$$

where

$$\varepsilon_m = \begin{cases} 1, & \text{if } m \text{ is even,} \\ 0, & \text{if } m \text{ is odd.} \end{cases}$$

On successively finding α_{2m} from (2.16), we discover that $\alpha_{2m}(p) = (-1)^{m+1} c_m p^m$, where c_m are positive constants, with $c_{m+1} < c_m$ for all $m = 1, 2, \dots$. Henceforth, the following assumptions will be made regarding the solution of the boundary value problem (2.10)–(2.13):

1) $u(x, t)$ has derivatives up to the k -th order with respect to t , and in addition, $\partial^s u / \partial t^s = O(1)$ as $x \rightarrow 0$, for any $t > 0, s = 0, 1, \dots, k$;

2) all these derivatives, regarded as functions of t , are function-originals.

Under these assumptions, on taking the inverse Laplace transformation in the relation

$$x \frac{dv}{dx} = v \sum_{m=1}^k (-1)^{m+1} c_m p^m x^{2m} + O(x^{2k+2}) v(x, p)$$

and noting that

$$\left. \frac{\partial^s u}{\partial t^s} \right|_{t \rightarrow 0+0} = L^s u|_{t \rightarrow 0+0} = 0, \quad s = 1, 2, \dots, k,$$

we find that

$$\frac{\partial u}{\partial x} = \sum_{m=1}^k (-1)^{m+1} c_m \frac{\partial^m u}{\partial t^m} x^{2m-1} + O(x^{2k+1}).$$

If conditions 1) and 2) hold for all $k = 0, 1, \dots$, then

$$\frac{\partial u}{\partial x} = \sum_{m=1}^{\infty} (-1)^{m+1} c_m \frac{\partial^m u}{\partial t^m} x^{2m-1}. \quad (2.17)$$

Since, in an alternating asymptotic series, the greatest accuracy is obtained when the series is broken off at the term which precedes the term of least absolute value, it follows from (2.17) that

$$\frac{\partial u}{\partial x} = \frac{x}{2} \frac{\partial u}{\partial t} + R_2(x, t),$$

where

$$R_2(x, t) = \sum_{h=2}^{\infty} (-1)^{h+1} c_h \frac{\partial^h u}{\partial t^h} x^{2h-1},$$

$$|R_2(x, t)| \leq \frac{1}{16} \left| \frac{\partial^2 u}{\partial t^2} \right| x^3 = O(x^3),$$

since the absolute value of the error R_2 is not greater than the absolute value of the first of the discarded terms. It is thus sufficient to require that conditions 1) and 2) hold only for $s = 0, 1, 2$, since further assumptions about the smoothness of $u(x, t)$ with respect to t do not improve the asymptotic convergence of the series (2.17). The behaviour of the manifold $S_{u, x, t} = \{u(x, t) = O(1), x < \varepsilon, t > 0\}$ of solutions of (2.10), (2.13), bounded as $x \rightarrow 0$, for any $t > 0$, in the x, t plane is thus described by the first-order partial differential equation

$$\frac{\partial u}{\partial x} = \frac{x}{2} \frac{\partial u}{\partial t} \quad (2.18)$$

up to an accuracy of ε^3 . Notice that, from (2.18) as $x \rightarrow 0$, we have $(\partial u / \partial x)|_{x=0} = 0$, as might be expected [5]; in addition, the relation

$$\frac{u(t, h) - u(t, 0)}{h} = \frac{h}{4} \frac{\partial u}{\partial t}(t, 0),$$

which approximates the condition $(\partial u / \partial x)|_{x=0} = 0$ on the solution of (2.10) to order h^2 , is easily obtained from (2.18). On writing the equation of the manifold (2.18) of bounded solutions, at a point $x_0 = h/2$, sufficiently close to the singular point, and noting that

$$\frac{\partial u}{\partial x} \Big|_{x_0} - \frac{u(t, h) - u(t, 0)}{h} = O(h^2), \quad u(t, x_0) - u(t, h) = O(h),$$

we get

$$\left(\frac{\partial u}{\partial x} - \frac{x}{2} \frac{\partial u}{\partial t} \right) \Big|_{x=x_0} - \left(\frac{u(t, h) - u(t, 0)}{h} - \frac{h}{4} \frac{\partial u(t, 0)}{\partial t} \right) = O(h^2),$$

which is in agreement with [5].

Note. The operator $\mathcal{L}_2[\partial/\partial x]$ in Eq. (1.1) can have the more general form

$$\mathcal{L}_2 \left[\frac{\partial}{\partial x} \right] = \frac{\partial^2}{\partial x^2} + B(x) \frac{\partial}{\partial x} + C(x),$$

where $B(x)$ and $C(x)$ are matrices of order q , which can simultaneously have irregular singularities of the 1st and 2nd kinds of integral order at infinity, and at the point $x = 0$, a regular singularity of the same integral order. In this case also, we can "move" in a similar way from the singular point $x = 0$ to a sufficiently close point x_0 , and from infinity to a sufficiently remote point x_∞ . After "displacement" of the entire linear manifold of bounded solutions to these points, the equation has to be approximated to the same order of accuracy as that possessed by the difference scheme of initial equations.

3. Comparison with solutions obtained by the method of straight lines

The displacement of the condition for boundedness from a regular and an irregular point for Eqs. (1.1)–(1.3) to the corresponding points of numerical integration, can also be performed by the method of straight lines. In order to compare the first approach with the second, we shall apply the method of straight lines e.g., to the boundary value problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad t > 0, \quad x \in (0, \infty), \quad u|_{t=0} = 0, \quad u|_{x=0} = q(t), \quad (3.1)$$

$$|u(x, t)| \rightarrow 0 \quad \text{as} \quad x \rightarrow \infty \quad \text{for any} \quad t > 0. \quad (3.2)$$

After discretization with respect to the time variable in the finite interval $[0, T]$ in (3.1), we obtain a system of ordinary differential equations, containing the small parameter τ :

$$\frac{u^{n+1} - u^n}{\tau} = \frac{d^2 u^{n+1}}{dx^2}, \quad \text{where} \quad \tau = \frac{T}{N}, \quad n = 0, 1, \dots, N,$$

with the condition that the solutions be bounded as $x \rightarrow \infty$. Introducing the vector $Y = \{u_1, \dots, u_N\}$, we rewrite (3.3) in the matrix form

$$d^2 Y / dx^2 + BY = 0, \quad (3.4)$$

where B is a constant N -th order matrix of the form

$$\frac{1}{\tau} \begin{vmatrix} -1 & & & & 0 \\ & 1 & -1 & & \\ & & 1 & -1 & \\ & & & \dots & \\ 0 & & & & 1 & -1 \end{vmatrix}.$$

The unknown matrix α , appearing in the equation of the manifold of solutions, stable as $x \rightarrow \infty$, of Eq. (3.4), which latter has an irregular singularity at infinity, is found from the relations

$$\alpha^4 + \alpha^2 + B = 0, \quad \alpha|_{x \rightarrow \infty} = \alpha_0 = -(-B)^{1/2},$$

$$\alpha \sim \sum_{k=0}^{\infty} \frac{\alpha_k}{x^k} \quad \text{as} \quad x \rightarrow \infty.$$

It is easily shown that all the eigenvalues of the matrix $\alpha_0 = -(-B)^{1/2}$ are equal to $-(1/\tau)^{1/2}$; hence the coefficients in the asymptotic expansion of the matrix α in the neighbourhood of infinity, depend on the small parameter τ ; this dependence is of the type $\alpha_k = O(\tau^{-k/2})$.

On refining the mesh with respect to t , the numbers α_k become infinitely large. Hence the point x_∞ , to which the boundedness condition is "displaced" from infinity, depends additionally on the step τ ; if $\tau' < \tau$, then $x_\infty(\tau) < x_\infty(\tau')$.

Condition (3.2) for Eq. (3.1) is thus approximated non-uniformly with respect to τ by the method of straight lines. To a first approximation we have

$$\left. \frac{dY}{dx} \right|_{x=x_\infty(\tau)} = -\tau^{-1/2} EY|_{x=x_\infty(\tau)}. \quad (3.5)$$

In other words,

$$\left. \frac{du^n}{dx} \right|_{x=x_\infty(\tau)} = -\tau^{-1/2} u^n|_{x=x_\infty(\tau)}, \quad n=1, 2, \dots, N.$$

The approach described in Section 1 gives, for sufficiently large x_∞ , the following integral approximation, which takes account of all the preceding layers:

$$\int_0^t \frac{\partial u(x_\infty, \sigma)}{\partial x} \frac{d\sigma}{[\pi(t-\sigma)]^{1/2}} = -u(x_\infty, t). \quad (3.6)$$

Notice that, as $x \rightarrow \infty$, the equation of the manifold of bounded solutions is the same, to a first approximation, for equation (3.3) as for the equation $u^{n+1} = \tau d^2 u^{n+1} / dx^2$. Consequently, on displaying the boundedness condition from layer to layer for Eq. (3.3) by the method of straight lines, the second $-u^n/\tau$ on the left-hand side of (3.3) is "frozen" for sufficiently large x , with the result that relation (3.5) becomes diagonal, as distinct from (3.6); this indicates one of the qualitative differences between the two approaches.

4. Application to a physical problem

The propagation of stationary axisymmetric light beams in a cubic medium is modelled in the parabolic equation approximation by the following non-linear boundary value problem [6, 7]:

$$i \frac{\partial u}{\partial z} = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + |u|^2 u, \quad 0 < z, r < \infty, \quad (4.1)$$

$$u|_{z=0} = \varphi(r), \quad (4.2)$$

$$\left. \frac{\partial u}{\partial r} \right|_{r=0} = 0, \quad (4.3)$$

$$|u| \rightarrow 0 \quad \text{as} \quad r \rightarrow \infty \quad \text{for any} \quad z > 0. \quad (4.4)$$

It is easily shown by direct calculation that Eq. (4.1) has the integral of motion

$$\int_0^\infty |u|^2 r \, dr = \text{const.} \quad (4.5)$$

In the light of (4.5), it is natural to assume from physical considerations that $u(r, z)$, as a function of r , decreases more rapidly than $1/r$ at infinity.

Hence the non-linear term in (4.1) decreases at infinity at the rate at least of $1/r^3$, so that it can be neglected for sufficiently large r . In view of this, the non-linear equation (4.1) can be replaced for large r by a linear equation, and the method described in Section 1 becomes applicable to the boundary value problem. On successively applying the procedure described in Section 1, we find after calculations that the equation of the manifold of solutions, bounded as $r \rightarrow \infty$, of the problem (4.1)–(4.4), is

$$\begin{aligned} & \frac{(2/\pi)^{1/2}}{1+i} \int_0^z \frac{\partial u(r, \xi)}{\partial r} \frac{d\xi}{(z-\xi)^{1/2}} = -u(r, z) \\ & + \sum_{k=1}^{\infty} \frac{(-1)^k c_k 2^{k/2}}{(1+i)^k} r^{-k} \int_0^z u(r, \xi) (z-\xi)^{(k-2)/2} d\xi - \varphi_0 \\ & + i \sum_{k=1}^{\infty} r^{-k} \sum_{j=0}^{\infty} \frac{(-1)^{k+2-j} c_{kj} \varphi_j}{(1+i)^{(k+2-j)}} \frac{2^{(k+2-j)/2}}{\Gamma((k+2-j)/2)} z^{k-j}, \end{aligned}$$

where $\Gamma(k/2)$ is the gamma function, φ_j , $j=0, 1, \dots$, are the coefficients of the asymptotic expansion of the function $\phi(r)$ close to infinity, $c_k = (1/2, 1/8, 1/8, 25/128, \dots)$, and c_{kj} is the triangular matrix

$$\begin{vmatrix} 1/2 & & 0 \\ 1/2 & -1 & \\ 1/8 & -1/2 & 1 \\ \vdots & \vdots & \vdots & \ddots \end{vmatrix}.$$

To a first approximation we have the relation

$$\frac{(2/\pi)^{1/2}}{1+i} \int_0^z \frac{\partial u(r_\infty, \xi)}{\partial r} \frac{d\xi}{(z-\xi)^{1/2}} = -u(r_\infty, z) - \varphi_0, \quad (4.6)$$

which correctly approximates the boundary condition (4.4) for Eq. (4.1). For the difference analogue of the boundary condition (4.6), we propose the following approximation with respect to z :

$$\begin{aligned} & \int_0^z \frac{\partial u(r_\infty, \xi)}{\partial r} \frac{d\xi}{(z-\xi)^{1/2}} = \sum_{k=0}^{n-1} \int_{z_k}^{z_{k+1}} \frac{\partial u(r_\infty, \xi)}{\partial r} \frac{d\xi}{(z-\xi)^{1/2}} \\ & \approx \sum_{k=0}^{n-1} \frac{1}{2} \left[\frac{\partial u(r_\infty, z_{k+1})}{\partial r} + \frac{\partial u(r_\infty, z_k)}{\partial r} \right] \int_{z_k}^{z_{k+1}} \frac{d\xi}{(z-\xi)^{1/2}} \end{aligned}$$

$$= \sum_{k=0}^{n-1} \left[\frac{\partial u(r_{\infty}, z_{k+1})}{\partial r} + \frac{\partial u(r_{\infty}, z_k)}{\partial r} \right] [(z - z_{k+1})^{1/2} - (z - z_k)^{1/2}].$$

The proposed method was realized on the BESM-6 computer. As the initial approximation we took the Gaussian distribution $\varphi(r) = \exp(-r^2 / 2l^2)$, where l is the characteristic width of the initial pencil. For numerical integration of problem (4.1)–(4.4), Eq. (4.1) was linearized and approximated by an implicit two-layer second-order finite difference scheme with respect to both coordinates. In order to allow for the singularity of the behaviour of the field close to the axis, and to take a sufficiently large (with respect to r) interval of numerical integration, a non-uniform r mesh was used, with a specific law of variation of the integration step. Condition (4.4) was approximated by relation (4.6). To check the accuracy of the computations, the conservation of the energy integral (4.5) was utilized. The results obtained are in agreement with those obtained in [6].

In conclusion, the author sincerely thanks A. A. Abramov for suggesting the problem and for his assistance, and also S. A. Gabov for valuable comments.

Translated by D. E. Brown

REFERENCES

1. BIRGER, E. S., and LYLIKOVA, N. B., On finding the solutions with a given condition at infinity for certain systems of ordinary differential equations, *Zh. vychisl. Mat. mat. Fiz.*, 5, No. 6, 979–990, 1965.
2. CODDINGTON, E. A., and LEVINSON, N., *Theory of ordinary differential equations*, McGraw, 1955.
3. BELLMAN, R., *Stability theory of differential equations*, Dover 1953.
4. ABRAMOV, A. A., Displacement of the boundedness condition for some systems of ordinary differential equations, *Zh. vychisl. Mat. mat. Fiz.*, 1, No. 4, 733–737, 1961.
5. SAMARSKII, A. A., *Introduction to the theory of difference schemes* (Vvedenie v teoriyu raznostnykh skhem), Nauka, Moscow, 1971.
6. DYSHKO, A. L., LUGOVOI, V. N., and PROKHOROV, A. M., Self-focusing of intense light pencils, *ZhETF Letters*, 6, No. 5, 655–659, 1967.
7. KELLEY, P. L., Self-focusing of optical beams, *Phys. Rev. Letters*, 15, 1005–1008, 1965.

ON RATIONAL APPROXIMATION IN CASE'S METHOD*

V. P. GORELOV and V. I. IL'IN

Moscow

(Received 12 August 1974; revised 4 March 1975)

AN APPROXIMATE description of the non-diffusion term in the neutron flux expression is given, and is justified numerically using elementary examples.

Introduction

It is well known that, in practical applications of neutron transport theory, a detailed knowledge of the space behaviour of the neutron flux is required, as well as high accuracy in computing the integral characteristics of the system. There are methods (see e.g., [1]) which provide highly accurate computation of integral characteristics such as the critical dimensions, yet do not yield information about the dependence of the flux on the space coordinate close to the interface (in the method of [1], use is only made of the fact that a jump is present in the asymptotic currents at the interface). The same feature is to be found in the zero approximation of the scheme developed in [3, 4] for solving multi-layer problems with the aid of Case's method [2]. Generally speaking, the schemes for solving such problems with the aid of Case's method (see [3, 4], and also [5, 6]) could enable the integral characteristics as well as the flux behaviour close to the interface to be computed to high accuracy. But it is actually very laborious to obtain higher than the zero approximations for the scheme of solution e.g., of [3].

In this connection, it seems sensible to try to describe the non-diffusion term in the flux expression in an elementary (rational) way, such that, on the one hand, good accuracy in computing the integral characteristics is achieved, and on the other hand, transition effects close to an interface can be approximately described. The possibility of such an approach, whereby the solution of singular integral equations can be avoided, is mentioned in [7]. The realization of a similar approach is referred to in [8] for the case of non-centralized layers; here, direct use is made of the scheme of solution of the two-zone problem with the aid of Case's method [2]. In the present paper we justify numerically a simple description of the non-diffusion term in the neutron flux expression, aimed at preserving the singularity in the behaviour of the solution close to an interface. Sections 1 and 2 deal with the plane and the spherical geometries respectively. In Section 3 we comment on the limits within which our proposed method may be used.

1. Milne's problem; critical dimension of the plate

The numerical justification of our description of the non-diffusion component of the neutron flux in a plane geometry will first be given for the case of Milne's problem for a non-absorbent medium (see e.g., [9]). As a preliminary, notice that, in the plane geometry, the angular dependence of the neutron flux $\Psi(x, \mu)$, where μ is the cosine of the angle between the direction of neutron

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 136-148, 1977.

movement and the positive x axis, has a discontinuity at the interface for $\mu = 0$. It is easily seen from simple geometry that

$$\Psi(x_0, -0) - \Psi(x_0, +0) = \frac{c_1 - c_2}{2} \Psi(x_0),$$

where c_1, c_2 are the numbers of secondary neutrons in collision to the right and left respectively of the interface $x = x_0$, and

$$\Psi(x) = \int_{-1}^1 \Psi(x, \mu) d\mu$$

is the neutron flux.

We shall show that, for this reason, $d\Psi(x)/dx$ is unbounded at $x = x_0$. From Boltzmann's equation one can obtain

$$\lim_{\delta \rightarrow 0} \left[\frac{d}{dx} \Psi(x) \right]_{x_0+\delta} = \lim_{\delta \rightarrow 0} \int_0^1 \frac{\Psi(x_0+\delta, -\mu) - \Psi(x_0+\delta, \mu)}{\mu} d\mu,$$

after which, on taking account of the discontinuity of $\Psi(x_0, \mu)$ for $\mu = 0$, it is easily seen that the derivative $d\Psi(x)/dx$ for $x = x_0$ is unbounded.

When describing approximately the non-diffusion component of the flux, we shall try to preserve the property just mentioned. For the flux in Milne's problem, in accordance with [2], we can write ($x \geq 0$)

$$\Psi(x, \mu) = H_1 + (x - \mu) + \int_0^1 e^{-x/v} H(v) \Phi(v, \mu) dv, \quad (1.1)$$

where H_1 and $(x - \mu)$ are the eigenfunctions of the discrete part of the spectrum $v \notin [-1, 1]$, describing the asymptotic (remote from the interface) behaviour of the neutron flux;

$$\Phi(v, \mu) = \frac{cv}{2} \frac{1}{v - \mu} + \lambda(v) \delta(v - \mu)$$

are the eigenfunctions of the continuous part of the spectrum $v \in (-1, 1)$, the expansion with respect to which in fact describes the non-diffusion component of the flux (1.1); $c = 1$,

$$\lambda(v) = 1 - \frac{v}{2} \ln \left(\frac{1+v}{1-v} \right);$$

$\delta(v)$ is the delta function, and x is measured in free path lengths.

If we solve strictly the problem in question, using the theory of singular integral equations (see e.g., [2]), we can show that $H(1) = 0$. We shall seek $H(v)$ approximately in the form

$$H(v) = (1-v)H_2. \quad (1.2)$$

This form of $H(v)$ ensures that $d\Psi(x)/dx$ is divergent at the interface $x = 0$. The approximation $H_2 = 0$ corresponds to an asymptotic consideration. The approximation (1.2) enables us to take

effective account in (1.1) of the non-diffusion term, describing the influence of the interface with the vacuum on the flux behaviour. To find H_1 and H_2 we use the Marshak boundary conditions [9]

$$\int_0^1 \mu^{2k+1} \Psi(0, \mu) d\mu = 0, \quad (1.3)$$

where $k = 0, 1$.

Substituting expression (1.1) in (1.3) and using (1.2), we obtain the following system of equations for H_1, H_2 :

$$wH = f,$$

where

$$\begin{aligned} f_k &= \frac{1}{2k+3}, \quad w_{k1} = \frac{1}{2k+2}, \\ w_{k2} &= \frac{1}{(2k+2)(2k+3)} - \frac{1}{2} \sum_{l=1}^{2k+1} \frac{1}{(2k+2-l)(2+l)(l+1)} \\ &\quad - \frac{1}{2} I_{2k+2}^{(+)}(0), \end{aligned}$$

and we use the notation

$$I_n^{(\pm)}(x) = \int_0^1 v^n (1-v) \ln \left(\frac{1}{v} \pm 1 \right) e^{-x/v} dv.$$

The $I_{2,4}^{(+)}(0)$ required for our computations are equal to $I_2^{(+)}(0) = 0.0871$, $I_4^{(+)}(0) = 0.03004$.

Of the two unknown coefficients, H_1 means physically the distance at which the asymptotic density of the neutrons vanishes when it is extrapolated into the vacuum [9]. Our computed value $H_1 = 0.71199$ differs from the exact value $H_T = 0.71045$ [9] by 0.2% (relative deviation). Notice that, in the asymptotic approximation, we have $H_1 = 0.66667$ (relative deviation 6%), while in the P_7 approximation of the method of spherical harmonics, we have $H_1 = 0.70692$ (relative deviation 0.5%) [10].

In addition, we computed:

1) the neutron flux at the vacuum interface, where the non-diffusion effect is a maximum:

$$\Psi(0) = \Psi(x)|_{x=0} = \frac{2}{J} \left[H_1 + x + \frac{H_2}{2} (E_2(x) - E_3(x)) \right] \Big|_{x=0},$$

$$E_n(x) = \int_1^\infty y^{-n} e^{-xy} dy,$$

where the quantity J normalizes the expression (1.1) in such a way that

$$\int_{-1}^0 \mu \Psi(0, \mu) d\mu = -1,$$

and is equal to $J = 0.5H_1 + 0.3 + 0.5H_2 \left[\frac{1}{6} - I_2^{(+)}(0) \right]$, while $H_1 = 0.71199$, $H_2 = -0.56957$, so that $J = 0.66666$;

2) the angular distribution of the neutrons leaving the half-space:

$$\Psi(0, -|\mu|) = \frac{1}{J\Psi(0)} \{H_1 + |\mu| + 0.5H_2[0.5 + |\mu| - |\mu| \ln(1 + 1/|\mu|)]\}.$$

The results obtained for $\Psi(0, -|\mu|)$ are given in Table 1.

TABLE 1

$ \mu $	$\Psi_0(0, - \mu)$	$\Psi(0, - \mu)$	$\Psi_\tau(0, - \mu)$	$\tilde{\Psi}(0, - \mu)$
0	0.5000	0.5000	0.5000	0.4936
0.1	0.5707	0.6287	0.6236	0.6207
0.2	0.6414	0.7330	0.7252	0.7238
0.3	0.7573	0.8313	0.8213	0.8207
0.4	0.7828	0.9265	0.9146	0.9148
0.5	0.8535	1.0199	1.0064	1.0070
0.6	0.9242	1.1121	1.0971	1.0980
0.7	0.9950	1.2035	1.1870	1.1880
0.8	1.0657	1.2941	1.2764	1.2777
0.9	1.1364	1.3845	1.3653	1.3668
1.0	1.2071	1.4744	1.4539	1.4560

Let us next indicate how the quantity $\Psi(0, -|\mu|)$ just determined can be improved in accuracy. It follows from simple geometry that, for a non-absorbent half-space we have

$$\tilde{\Psi}(x, \mu) = \begin{cases} 0.5 \int_0^{x/\mu} e^{-\xi} \Psi(x - \xi\mu) d\xi, & \mu > 0, \\ 0.5 \int_0^{\infty} e^{-\xi} \Psi(x - \xi\mu) d\xi, & \mu < 0. \end{cases}$$

If we substitute the $\Psi(x)$ defined above in this relation, we can obtain for $\tilde{\Psi}(0, -|\mu|)$ the expression

$$\tilde{\Psi}(0, -|\mu|) = \frac{1}{J\tilde{\Psi}(0)} \left\{ H_1 + |\mu| + 0.5H_2 \left[0.5 + |\mu| - |\mu| (1 + |\mu|) \ln \left(1 + \frac{1}{|\mu|} \right) \right] \right\},$$

where H_1, H_2, J are given above. The quantity $\tilde{\Psi}(0) = 1.731$ deviates from the exact value $\Psi_\tau(0) = 1.732$ (see [11]) by 0.06%, whereas $\Psi(0) = 1.709$ has a relative deviation of 1.3%. The values of the function $\tilde{\Psi}(0, -|\mu|)$ are also quoted in Table 1. The function $\Psi_0(0, -|\mu|)$, also in Table 1, corresponds to the asymptotic approximation ($H_2 = 0$), while $\Psi_\tau(0, -|\mu|)$ is computed to high accuracy in [12].

To sum up, the use of the approximation (1.2) has provided satisfactory computational accuracy both for the integral characteristic H_1 , and for the behaviour of the neutron flux close to the interface.

Now consider the problem of finding the critical half-thickness of a plate of neutron multiplying material. Using the method described in [2] and symmetry arguments (the interface $x = \pm a$), we can write for the neutron flux:

$$\begin{aligned} \Psi(x, \mu) = & H_1 \frac{v_1 \cos(x/v_1) + \mu \sin(x/v_1)}{v_1^2 + \mu^2} + H_1(|\mu|) \lambda_1(\mu) e^{\mp x/|\mu|} \\ & + \frac{c_1}{2} \int_0^1 v H_1(v) \left(\frac{e^{-x/v}}{v-\mu} + \frac{e^{x/v}}{v+\mu} \right) dv, \\ \lambda_1(\mu) = & 1 - \frac{c_1 \mu}{2} \ln \left(\frac{1+\mu}{1-\mu} \right); \end{aligned} \quad (1.4)$$

the upper sign refers to $\mu > 0$, and the lower to $\mu < 0$.

We shall seek $H_1(v)$ as

$$H_1(v) = (1 - |v|) H_2 e^{-a/|v|}. \quad (1.5)$$

It can be shown that the exponential factor also appears when the problem is considered strictly. Expression (1.5) provides divergence of $d\Psi(x)/dx$ for $x = \pm a$. We find the constants H_1 and H_2 by using Marshak's conditions [9]:

$$\int_0^{-1} \mu^{2k+1} \Psi(a, \mu) d\mu = 0, \quad (1.6)$$

where $k = 0, 1$.

After substituting expressions (1.4) in (1.6) and using (1.5), we obtain the following system of homogeneous equations in H_1, H_2 :

$$wH = 0, \quad (1.7)$$

where

$$\begin{aligned} w_{11} = & \frac{v_1}{2} \ln(1+v_1^{-2}) \cos \frac{a}{v_1} - \frac{c_1-1}{c_1} \sin \frac{a}{v_1}, \\ w_{21} = & \frac{v_1}{2} [1 - v_1^2 \ln(1+v_1^{-2})] \cos \frac{a}{v_1} - \left[\frac{1}{3} - v_1^2 \frac{c_1-v_1}{c_1} \right] \sin \frac{a}{v_1}, \\ w_{12} = & \frac{c_1}{2} \left[\mathcal{E}_1(2a) - \mathcal{E}_1(0) - I_2^{(+)}(2a) - I_2^{(+)}(0) + \frac{2}{c_1} \mathcal{E}_1(0) \right], \\ w_{22} = & \frac{c_1}{2} \left[\sum_{l=1}^3 (-1)^{l-1} \frac{\mathcal{E}_l(2a)}{4-l} - \sum_{l=1}^3 \frac{\mathcal{E}_l(0)}{4-l} + \right. \end{aligned}$$

$$+ \frac{2}{c_1} \mathcal{E}_3(0) - I_4^{(+)}(2a) - I_4^{(+)}(0) \Big],$$

$$\mathcal{E}_n(x) = \int_0^1 v^n (1-v) e^{-x/v} dv = E_{n+2}(x) - E_{n+3}(x).$$

The critical half-thickness of the plate a is the least positive root of the equation

$$\operatorname{tg} \frac{a}{v_1} = \frac{c_1 v_1 \ln(1+v_1^{-2})}{2(c_1-1)} \left\{ 1 - \frac{[\ln(1+v_1^{-2})]^{-1} - c_1/3(c_1-1)}{w_{22}/w_{12} + v_1^2 - c_1/3(c_1-1)} \right\}.$$

Computational results from this last expression are given in Table 2, where they are compared with the results obtained in the S_{16} approximation by Carlson's method, and in the V_2 approximation of the variational method (see [13]). The values of a_0 correspond to the asymptotic approximation, and in the parentheses we quote the relative deviations from the values a_{V_2} (in %), which represent the most accurate values. The quantities v_1^{-1} were borrowed from [1], where they were computed for a large set of c_1 values.

TABLE 2

	c_1			
	1.05	1.10	1.20	1.40
a_0	3.3344 (1.03)	2.1388 (1.20)	1.3014 (0.94)	0.7322 (0.60)
a	3.2989 (0.04)	2.1124 (0.05)	1.2896 (0.02)	0.7394 (0.40)
a_{V_2}	3.3002	2.1134	1.2893	0.7366
$a_{S_{16}}$	3.3023 (0.06)	2.1146 (0.05)	1.2902 (0.06)	0.7372 (0.10)

2. Critical size of the sphere

We shall first consider the problem of finding the critical radius of a sphere of homogeneous, neutron multiplying material. We know from [1, 9] that the Peierls' integral equation can in this case be reduced to an integral equation, formally identical with the equation for the plane geometry, which corresponds to an integro-differential equation, of the same form as Boltzmann's equation in the x geometry. The solution of this equation, i.e., the pseudo-distribution $\Psi(r, u)$, is connected with the true flux $\Psi(r)$ by the relation

$$r\Psi(r) = \int_{-1}^1 \Psi(r, u) du$$

(here, $u \in [-1, 1]$ is a non-physical parameter).

On the outer surface, $\Psi(r, u)$ has to satisfy the condition

$$\Psi(R, u < 0) = 0, \quad (2.1)$$

and moreover,

$$\Psi(r, u) = -\Psi(-r, -u). \quad (2.2)$$

Here, R is the required radius of the sphere, expressed in free path lengths.

The method described in [2] can be used to find $\Psi(r, u)$; hence, using (2.2), we write

$$\begin{aligned} \Psi_1(r, u) = & H_1^{(4)} \frac{v_1 \sin(r/v_1) - u \cos(r/v_1)}{v_1^2 + u^2} \pm H_1(|u|) \lambda_1(u) e^{\mp r/|u|} \\ & + \frac{c_1}{2} \int_0^1 v H_1(v) \left(\frac{e^{-r/v}}{v-u} - \frac{e^{r/v}}{v+u} \right) dv, \end{aligned} \quad (2.3)$$

where the upper sign refers to $u > 0$, and the lower to $u < 0$, and the lower to $u < 0$. If we solve problem (2.1) strictly for (2.3), we can show that $H_1(1) = 0$ and $H_1(|u|) \sim e^{-R/|u|}$. We shall seek $H_1(v)$ in the form

$$H_1(|v|) = (1 - |v|) e^{-R/|v|} H_2^{(4)}. \quad (2.4)$$

This approximation implies that $d\Psi(r)/dr$ is divergent for $r = R$. The latter property also follows from the transport equation in spherical geometry. From it we can obtain directly, for the homogeneous sphere,

$$\begin{aligned} \lim_{\delta \rightarrow 0} \left[\frac{d}{dr} \Psi(r) \right]_{R+\delta} = & \lim_{\delta \rightarrow 0} \left\{ \int_0^1 \frac{\Psi(R+\delta, -\mu) - \Psi(R+\delta, \mu)}{\mu} d\mu - \right. \\ & \left. - \frac{1}{R+\delta} \int_0^1 \frac{(1-\mu^2)}{\mu} \left[\frac{\partial}{\partial \mu} \Psi(R+\delta, \mu) + \frac{\partial}{\partial \mu} \Psi(R+\delta, -\mu) \right] d\mu \right\} \end{aligned}$$

(μ is the cosine of the angle between the neutron velocity vector and the radius vector R). Simple geometry shows that, if $\Psi(R, -0) = \Psi(R, +0)$, then the derivative $\partial \Psi(R, \mu)/\partial \mu$ has a discontinuity for $\mu = 0$ (see also the computations in [14]). Hence it follows at once that $d\Psi(r)/dr$ is unbounded on the interface with the vacuum.

The unknown coefficients in (2.3) may be found by replacing the exact condition (2.1) by approximate conditions, representing an analogue of Marshak's conditions:

$$\int_0^{-1} u^{2k+1} \Psi_1(R, u) du = 0, \quad (2.5)$$

where $k = 0, 1$.

Substituting Eq. (2.3) into Eq. (2.5) and using Eq. (2.4), we obtain a system of homogeneous equations of the type (1.7), where

$$w_{11} = \frac{v_1}{2} \ln(1 + v_1^{-2}) \sin \frac{R}{v_1} + \frac{c_1 - 1}{c_1} \cos \frac{R}{v_1},$$

$$w_{21} = \frac{v_1}{2} [1 - v_1^2 \ln(1 + v_1^{-2})] \sin \frac{R}{v_1} + \left[\frac{1}{3} - v_1^2 \frac{(c_1 - v_1)}{c_1} \right] \cos \frac{R}{v_1},$$

$$w_{12} = \frac{c_1}{2} \left[\mathcal{E}_1(2R) + \mathcal{E}_1(0) + I_2^{(+)}(0) - I_2^{(+)}(2R) - \frac{2}{c_1} \mathcal{E}_1(0) \right],$$

$$w_{22} = \frac{c_1}{2} \left[\sum_{l=1}^3 \frac{(-1)^{l-1} \mathcal{E}_l(2R) - \mathcal{E}_l(0)}{4-l} + I_4^{(+)}(0) - I_4^{(+)}(2R) - \frac{2}{c_1} \mathcal{E}_3(0) \right]$$

and all the notation is the same as in the previous section.

The condition for solvability of system (2.5) is

$$\operatorname{tg} \frac{R}{v_1} = - \frac{2(c_1 - 1)}{c_1 v_1 \ln(1 + v_1^{-2})} \frac{v_1^2 + w_{22}/w_{12} - c_1/3(c_1 - 1)}{v_1^2 + w_{22}/w_{12} - [\ln(1 + v_1^{-2})]^{-1}}, \quad (2.6)$$

and the required radius R is the least positive root of this equation. Notice that, just as in the computations of a in Table 2, we have neglected the dependence of R on the matrix elements w_{12} and w_{22} when evaluating R ; this considerably simplifies the working. The results are quoted in Table 3, where they are compared with the exact values R_T of [13], and also with the values R_{S16} , computed in the S_{16} approximation (see [13]). The values $R_{\Gamma,T}$ in Table 3 (see [13]) refer to the

TABLE 3

Radii	c_1		
	1.05	1.20	1.40
R_0	7.3118 (0.50)	3.1839 (0.40)	1.9794 (0.30)
R	7.2764 (0.01)	3.1719 (0.001)	1.9866 (0.07)
R_T	7.2772	3.1720	1.9854
R_{S16}	7.2723 (0.07)	3.1690 (0.10)	1.9830 (0.14)
$R_{\Gamma,T}$	7.2772	3.1720	1.9853

improved diffusion method [1], while the R_0 values were computed in the asymptotic approximation ($H_2^{(1)} = 0$). We quote in parentheses the relative deviations (in %) from R_T .

The R_0 values were computed from the asymptotic formula

$$\operatorname{tg} \frac{R_0}{v_1} = - \frac{2(c_1 - 1)}{c_1 v_1 \ln(1 + v_1^{-2})}.$$

Now consider the problem of finding the critical radius of a sphere, surrounded by an infinite reflector. Given a constant free path length, the problem can again be reduced in this case (see [1, 9]) formally to the plane case. As before, the pseudo-distribution $\Psi(r, u)$ is antisymmetric (see (2.2)), it must be continuous at $r = R$, and it must vanish at infinity. For $0 < r < R$, it is described by the expression (2.3). For $R < r < \infty$, using Case's method, we can write

$$\begin{aligned}\Psi_2(r, u) &= H_1^{(2)} \frac{e^{-r/v_2}}{v_2 - u} + \frac{c_2}{2} \int_0^1 v H_2(v) \frac{e^{-r/v}}{v - u} dv \\ &+ \theta(u) H_2(|u|) \lambda_2(u) e^{-r/u}, \\ \theta(u) &= \begin{cases} 1, & u \geq 0, \\ 0, & u < 0. \end{cases}\end{aligned}\quad (2.7)$$

On considering the problem strictly, just as for example in [3], we can show that $H_1(|u|) \sim e^{-R/|u|}$, $H_2(|u|) \sim e^{-R/|u|}$ and $H_1(1) = H_2(1) = 0$. For an approximate description of $H_1(|v|)$ we shall use (2.4), while we write $H_2(|v|)$ as

$$H_2(|v|) = (1 - |v|) e^{R/|v|} H_2^{(2)}. \quad (2.4')$$

This form of $H_2(|v|)$ ensures that $d\Psi(r)/dr$ is divergent when approaching $r = R$ from the right.

TABLE 4

Radii	c_1		
	1.10	1.30	1.60
R_0	3.2943 (0.08)	1.6530 (1.04)	1.0357 (2.00)
R	3.2902 (0.06)	1.6730 (0.03)	1.0706 (0.01)
R_T	3.2923	1.6724	1.0705
$R_{T.T.}$	3.2940 (0.06)	1.6721 (0.02)	1.0697 (0.08)

We replace the exact continuity condition on the interface by the approximate condition

$$\int_0^{\pm 1} u^{2k+1} \Psi_1(R, u) du = \int_0^{\pm 1} u^{2k+1} \Psi_2(R, u) du, \quad (2.8)$$

where $k = 0, 1$. The choice (2.8) of the boundary conditions ensures that the asymptotic flux jumps at the interface, this being one of its known properties (see [1, 9]). After substituting (2.3) and (2.7) and using (2.4), (2.4') in (2.8), we obtain a system of homogeneous equations in $H_1^{(1)}$, $H_1^{(2)}$, $H_2^{(1)}$, $H_2^{(2)}$. The least positive root of its determinant is in fact the required critical radius R of the sphere with the infinite reflector. The computational results for R are given in Table 4; $c_2 = 0.99$. The values R_0 refer to the asymptotic approximation ($H_2^{(1)} = H_2^{(2)} = 0$); R_T is the value obtained in [6] for a large number of iterations in the strict scheme of solutions, using Case's method; $R_{T.T.}$ is the value computed by us from the improved diffusion theory of [1]; the relative deviations (in %) from R_T are quoted in parentheses.

The quantity R_0 was computed from the expression

$$\operatorname{tg} \frac{R_0}{v_1} = \frac{(c_1 - 1)}{c_1 v_1 \ln(1 + v_1^{-2})} \frac{c_2 v_2 \ln(1 - v_2^{-2})}{(1 - c_2)}. \quad (2.9)$$

When computing R , we neglected the dependence on R in expressions of the type

$$w_{2h+1} = \frac{c_1}{2} \left\{ \int_0^{\mp 1} u^{2h+1} du \left[\int_0^1 v(1-v) \left(\frac{e^{-2R/v}}{v-u} - \frac{1}{v+u} \right) dv \right. \right. \\ \left. \left. \pm \frac{2}{c_1} (1-|u|) e^{-R \pm R/|u|} \lambda_1(u) \right] \right\}.$$

Where more detailed computations are required, the functions $I_{24}^{(\pm)}(x)$, $\mathcal{E}_{123}(x)$ can easily be tabulated.

Let us now indicate some features of our approach in the case when the outer layers are finite, taking as our basis some of the conclusions of [8]. For clarity, we consider the two-zone reactor. Assume, as before, that the inner sphere consists of neutron multiplying material. The neutron pseudo-distribution in it will be described by expression (2.3) in the approximation (2.4). We can write in the shell ($r > 0$), using Case's method and symmetry considerations,

$$\Psi_2(r, u) = H_2^{(2)} \frac{v_2 \operatorname{sh}(r/v_2) - u \operatorname{ch}(r/v_2)}{v_2^2 - u^2} \\ + H_1^{(2)} \frac{v_2 \operatorname{ch}(r/v_2) + u \operatorname{sh}(r/v_2)}{v_2^2 - u^2} \\ + \frac{c_2}{2} \int_0^1 v \left[H_2^{(1)}(v) \frac{e^{-r/v}}{v-u} - H_2^{(2)}(v) \frac{e^{r/v}}{v+u} \right] dv \\ + [\theta(u) H_2^{(1)}(|u|) - \theta(-u) H_2^{(2)}(|u|)] \lambda_2(u) e^{-r/u}.$$

It can be shown in the same way as in [8] by a strict consideration of the problem that, in view of the continuity conditions on the interface, and the absence of $\Psi_2(R_0, u)$ for $u < 0$ we have $H_2^{(1,2)}(1) = 0$, $H_2^{(1)}(|u|) \sim e^{R/|u|}$, $H_2^{(2)}(|u|) \sim e^{-R_0/|u|}$, where R is the radius of the central sphere in free path lengths, and R_0 is the shell radius in the same units. These facts imply that the derivative $d\Psi(r)/dr$ has a singularity both for $r = R$ and for $r = R_0$. Hence we can seek $H_2^{(1,2)}(|v|)$ as

$$H_2^{(1)}(|v|) = (1-|v|) H_3^{(2)} e^{R/|v|}, \quad H_2^{(2)}(|v|) = (1-|v|) H_4^{(2)} e^{-R_0/|v|}.$$

The exact boundary conditions, for continuity of $\Psi(r, u)$ at $r = R$, and for $\Psi_2(R_0, u < 0) = 0$ are replaced by the conditions

$$\int_0^{-1} \Psi_2(R_0, u) u^{2h+1} du = 0, \\ \int_0^{\pm 1} \Psi_1(R, u) u^{2h+1} du = \int_0^{\pm 1} \Psi_2(R, u) u^{2h+1} du,$$

where $k = 0, 1$. This system of equations in fact enables us to find $H_2^{(1)}, H_{1234}^{(2)}$, while the condition for solvability of the system provides the equation for finding the critical dimension of the system.

3. The range of application. Conclusions

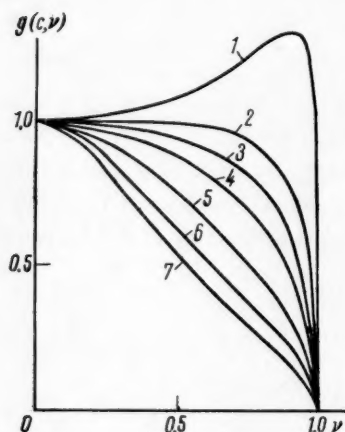


FIG. 1.

Curves: 1 — for $c=0.6$, 2 — for $c=0.8$, 3 — for $c=0.9$, 4 — for $c=0.99$, 5 — for $c=1.2$, 6 — for $c=1.4$, 7 — for $c=1.6$

TABLE 5

c_1/c_2	R_0	R_T [18]	R
1.1/0.8	4.4128 (0.73)	4.3809	4.3751 (0.13)
1.3/0.8	2.1105 (0.25)	2.1052	2.1020 (0.15)
1.6/0.8	1.2617 (1.08)	1.2755	1.2740 (0.12)
1.1/0.5	4.7352 (0.84)	4.6959	4.6886 (0.15)
1.3/0.5	2.3082 (0.53)	2.2962	2.2922 (0.17)
1.6/0.5	1.3798 (0.61)	1.3883	1.3862 (0.15)

Our description of the non-diffusion term in the neutron flux expression primarily utilizes the fact that the coefficient of the eigenfunction expansion of the continuous part of the spectrum ν in Case's method has a zero at $|\nu| = 1$. The reason for the vanishing of this function at $|\nu| = 1$, both when considering Milne's problem, and when considering critical problems, is that the following condition proves to hold:

$$H(\nu) \sim [\lambda^2(\nu) + \pi^2 c^2 \nu^2 / 4]^{-1} \\ = g(c, \nu).$$

In Fig. 1 we show curves of $g(c, \nu)$ for several values of c . It can be seen that, as c decreases (the absorption without neutron reproduction increases) a maximum occurs in $g(c, \nu)$ as a function of ν , i.e., the dependence becomes markedly non-linear. It is useful in this connection to set a limit to the range in which our method is applicable, dependent on the absorption in the shell. It turns out that, e.g., when computing the critical dimensions of spheres with a reflector, the approximation (2.4), (2.4') has a wide range of application. In Table 5 we give the critical radii R of the sphere with infinite reflector when there is substantial absorption in the latter, computed by the method described in Section 2. The R_0 values were found from (2.9). The R_T values are borrowed from [15], where they are obtained for a large number of iterations in the strict scheme of solution using Case's method [2]. The relative deviations (in %) from R_T are quoted in parentheses.

In short, our description of the non-diffusion component of the neutron flux enables the critical dimension to be computed with satisfactory accuracy. The accuracy is no worse than in the S_{16} approximation of Carlson's method or the improved diffusion method [1]. As compared with the former, it has the advantage of being less laborious, and as compared with the latter, it also enables the behaviour of the neutron flux close to the interface to be described. For instance, for the homogeneous sphere, the neutron flux is equal to

$$\Psi(r) = \Psi_{\text{ass}}(r) + \Psi_{\text{tr}}(r),$$

where

$$\Psi_{\text{ass}}(r) = \frac{2}{c_1 v_1} H_1^{(1)} \frac{\sin(r/v_1)}{r}, \quad \Psi_{\text{tr}}(r) = H_2^{(1)} \frac{\mathcal{E}_0(R+r) - \mathcal{E}_0(R-r)}{r};$$

the expression for $\mathcal{E}_n(x)$ is given above, while $H_2^{(1)}/H_1^{(1)} = -w_{11}/w_{12}$, w_{11} , w_{12} are given in Section 2, and R is the least positive root of Eq. (2.6).

Notice that $|\Psi_{\text{tr}}(r)|$ takes its maximum value on the interface with the vacuum $r = R$, and $\Psi_{\text{tr}}(0) = 0$. If it is necessary to refine the angle-space flux, the scheme described in Section 1 may be used; a simple illustrative example showed that this scheme is quite efficient.

Translated by D. E. Brown

REFERENCES

1. ROMANOV, YU. A., An improved diffusion method, in: *Studies of the critical parameters of reactor systems* (Issl. kritich. parametrov reaktornykh sistem), Atomizdat, 3-26, Moscow, 1960.
2. CASE, K. M., Elementary solutions of the transport equation and their applications, *Ann. Phys.*, **9**, 1-23, 1960.

3. GORELOV, V. P., and YUFEREV, V. I., Solution of the neutron transport equation in multi-layer plane and spherically symmetric systems, *Zh. vychisl. Mat. mat. Fiz.*, **11**, No. 1, 129–136, 1971.
4. GORELOV, V. P., IL'IN, V. I., and YUFEREV, V. I., Solution of some problems in neutron transport theory by the method of generalized eigenfunctions, *Zh. vychisl. Mat. mat. Fiz.*, **12**, No. 5, 1245–1264, 1972.
5. KUSZELL, A., The critical problem for multilayer slab systems, *Acta phys. polon.*, **20**, 567–589, 1961.
6. LEUTHAUSER, K.-D., Neutron transport in monoenergetic configurations with finite reflectors, *Atomkernenergie*, **14**, 148, 1969.
7. BLINKIN, V. L., and TROYANSKII, V. B., A generalized wave method for solving the neutron transport equation in the single-group approximation, in: *Physics of nuclear reactors* (Fiz. yadernykh reaktorov), No. 1, 167–177, Atomizdat, Moscow, 1968.
8. GORELOV, V. P., IL'IN, V. I., and POVYSHEV, V. M., On the probability of neutron absorption in plane and spherical blocks, *Zh. vychisl. Mat. mat. Fiz.*, **15**, No. 2, 391–403, 1975.
9. DAVISON, B., and SYKES, J. B., *Neutron transport theory*, Oxford U.P., 1957.
10. WEINBERG, A. M., and WIGNER, E. P., *Physical theory of neutron chain reactors*, U. of Chicago Press, 1958.
11. PTITSYNA, N. V., Use of the variational method in some generalizations of Milne's problem, in: *Some mathematical problems of neutron physics* (Nekotorye matem. zadachi neitronnoi fiz.), 28–55, Izd-vo MGU, Moscow, 1960.
12. PLACZEK, G., The angular distributions of neutrons emerging from a plane surface, *Phys. Rev.*, **72**, 556–558, 1947.
13. CARLSON, B., and BELL, J., Solution of the transport equation by the S_n method, in: *Physics of nuclear reactors* (Fiz. yadernykh reaktorov), 408–432, Atomizdat, Moscow, 1959.
14. NIKOLAISHVILI, SH. S., and LYASHENKO, E. I., Calculation of the angular distributions of neutrons in nuclear reactors, in: *Topics in reactor protection physics* (Vopr. fiz. zashchity reaktorov), No. 3, 4–68, Atomizdat, Moscow, 1969.
15. LEUTHAUSER, K.-D., Neutron transport in monoenergetic critical configurations with finite reflectors, Part 2, Spheres, *Atomkernenergie*, **14**, 254–256, 1969.

NON-LINEAR MATHEMATICAL PROBLEMS OF THE TRANSMISSION OF EXCITATORY AND INHIBITORY PULSES IN NERVE TISSUE*

S. F. MOROZOV and I. P. SMIRNOV

Gor'kii

(Received 30 June 1975)

FOR A non-linear integro-differential system of equations of the transmission of excitatory and inhibitory pulses in nerve tissues, in the standard and non-standard cases, existence and uniqueness theorems of the solution, prior estimates and some quantities of the solution are established.

Consideration of the single-velocity equations of the transmission of excitatory (ψ_+) and inhibitory (ψ_-) pulses in nerve tissues (see [1]) leads to a study of a non-linear integro-differential system of equations of the following form:

$$\begin{aligned} & \frac{\partial}{\partial t} \psi_{\pm}(s, P, t) + (s, \nabla) \psi_{\pm}(s, P, t) + \Sigma^{\pm}(s, P) \psi_{\pm}(s, P, t) \\ &= \int_{\Omega} W_{\pm}(s, s') \Sigma^{\pm}(s', P) \psi_{\pm}(s', P, t) ds' + Q^{\pm}(s, P) F_{\pm}(\psi_+, \psi_-), \end{aligned} \quad (1)$$

Zh. vychisl. Mat. mat. Fiz.*, **17, 1, 149–161, 1977.

$$\begin{aligned} & \frac{\partial}{\partial t} \psi_{-}(s, P, t) + (s, \nabla) \psi_{-}(s, P, t) + \Sigma^{-}(s, P) \psi_{-}(s, P, t) \\ &= \int_{\Omega} W_{-}(s \cdot s') \Sigma^{-}(s', P) \psi_{-}(s', P, t) ds' + Q^{-}(s, P) F_{-}(\psi_{+}, \psi_{-}), \end{aligned} \quad (\text{cont'd})$$

with the initial and boundary conditions

$$\psi_{\pm}(s, P, t) = \varphi_{\pm}'(s, P, t), \quad P \in \Gamma, \quad (n(P), s) < 0, \quad t \in [0, T], \quad (2)$$

$$\psi_{\pm}(s, P, t)|_{t=0} = \psi_{\pm}^{(0)}(s, P). \quad (3)$$

Here $P = \{x_1, x_2, x_3\} \in G$ is a convex domain in E^3 with the smooth boundary Γ , $n(P)$ is the outward normal at the point $P \in \Gamma$, $s = \{s_1, s_2, s_3\} \in \Omega$ is a unit sphere in E^3 , $t \in [0, T]$, $s \cdot s' = s_1 s_1' + s_2 s_2' + s_3 s_3'$.

The non-linear operators $F_{\pm}(\psi_{+}, \psi_{-})$ are considered in the MacCullagh-Pitts approximation (section 1).

In the present paper we study successively the stationary problem (section 2), for which existence and uniqueness theorems of the solution are established, prior estimates in terms of the data of the problem, and also some properties of the solution, following from additional conditions on the coefficients, and the non-stationary problem (section 3), for which theorems of the existence and uniqueness (and stability in L_2) of the solution are established. Section 3 is devoted to the definition of the fundamental spaces and operators of the problem and to an investigation of their properties.

1. Fundamental definitions

1. The coefficients. We assume that the functions $\Sigma^{\pm}(s, P)$, $Q^{\pm}(s, P)$ are measurable with respect to the ensemble of variables for $(s, P) \in \Omega \times G$ and satisfy almost everywhere in the following constraints:

$$\begin{aligned} 0 < \sigma^{\pm} \leq \Sigma^{\pm}(s, P) \leq \Sigma^{\pm} < \infty, \quad \sigma^{\pm}, \Sigma^{\pm} = \text{const}, \\ |Q^{\pm}(s, P)| \leq Q^{\pm}, \quad Q^{\pm} = \text{const}. \end{aligned}$$

The functions $W_{\pm}(s \cdot s')$ are integrable on $\Omega \times \Omega'$; $\xi_{\pm}(s, P)$, $\eta_{\pm}(s, P)$, $\mu_{\pm}(P, P')$, $s \in \Omega$, $P, P' \in G$, are coefficients occurring in the construction of the non-linear operators $F_{\pm}(\psi_{+}, \psi_{-})$ (see below), and measurable with respect to the ensemble of variables in the corresponding domains of definition.

2. The fundamental spaces. In what follows we use the Banach spaces \mathcal{H}_p^{\pm} of functions $\psi(s, P)$ measurable on $\Omega \times G$ with norms

$$\|\psi\|_p^{\pm} = \left[\int_{\Omega \times G} \Sigma^{\pm}(s, P) |\psi(s, P)|^p ds dP \right]^{1/p}$$

and their product $\mathcal{H}_p = \mathcal{H}_p^+ \times \mathcal{H}_p^-$ with norm $\|\psi\|_p = [(\|\psi_+\|_p^+)^p + (\|\psi_-\|_p^-)^p]^{1/p}$, $\psi = \text{col } \{\psi_+, \psi_-\}$, $1 < p < \infty$. The Banach spaces of functions $\varphi(s, P)$, $(s, P) \in \Omega \times G$, possessing the generalized derivative $(s, \nabla)\varphi \in \mathcal{H}_p^\pm$, with norms $\|\varphi\|_{W_p^\pm} = \|\varphi\|_p^\pm + \|(s, \nabla)\varphi\|_p^\pm$ will be denoted by W_p^\pm . We denote by D_p^\pm the subspaces of functions $\varphi(s, P) \in W_p^\pm$, satisfying the condition $\varphi(s, P) = 0$ for $P \in \Gamma_-^s$, where Γ_-^s is the part of the boundary Γ , on which $(s, n(P)) < 0$. Let $W_p = W_p^+ \times W_p^-$, $D_p = D_p^+ \times D_p^-$, D_p and W_p be everywhere dense in \mathcal{H}_p [2].

3. *The operators.* We define on \mathcal{H}_p the linear operator $S = \text{diag } \{S_+, S_-\}$, where

$$S_\pm \psi_\pm = [\Sigma^\pm(s, P)]^{-1} \int_\Omega W_\pm(s, s') \Sigma^\pm(s', P) \psi_\pm(s', P) ds', \quad \psi_\pm \in \mathcal{H}_p^\pm.$$

The operators generated on W_p^\pm, D_p^\pm by the linear differential expressions $l_\pm \varphi_\pm = [\Sigma^\pm(s, P)]^{-1} (s, \nabla) \varphi_\pm + \varphi_\pm$, will be denoted by \mathcal{L}_\pm and L_\pm respectively. On W_p and D_p respectively we introduce the operators $\mathcal{L} = \text{diag } \{\mathcal{L}_+, \mathcal{L}_-\}$, $L = \text{diag } \{L_+, L_-\}$. Finally we define the non-linear operators $F_\pm(\psi_+, \psi_-)$. For this purpose we introduce the linear operators $R_\pm: \mathcal{H}_p \rightarrow L_q(G)$, $1 \leq q \leq \infty$:

$$R_\pm \psi = \int_{\Omega \times G} [\xi_\pm(s', P') \psi_+(s', P') - \eta_\pm(s', P') \psi_-(s', P')] \times \mu_\pm(P, P') ds' dP', \quad \psi = \text{col } \{\psi_+, \psi_-\}.$$

Let $\mathcal{F}(x) = (u_0 + x)\theta(u_0 + x)$, $u_0 = \text{const}$, $x \in (-\infty, \infty)$, $\theta(x) = \{1 \text{ for } x \geq 0; 0 \text{ for } x < 0\}$. We define on $L_q(G)$ the non-linear superposition operator F by the equation

$$F\psi = \mathcal{F}(\psi(P)), \quad \psi(P) \in L_q(G), \quad 1 \leq q \leq \infty.$$

Then $F_\pm(\psi_+, \psi_-) = FR_\pm \psi$, $\psi = \text{col } \{\psi_+, \psi_-\} \in \mathcal{H}_p$.

4. *The properties of the operators.* The following lemmas hold.

Lemma 1 (see [2, 3]).

The operators $\mathcal{L}: W_p \rightarrow \mathcal{H}_p$, $L: D_p \rightarrow \mathcal{H}_p$ are bounded. The operator $L^{-1}: \mathcal{H}_p \rightarrow \mathcal{H}_p$ exists and

$$\|L^{-1}\|_p \leq \max \{1 - \exp(-\Sigma^+ d); 1 - \exp(-\Sigma^- d)\}, \\ d = \text{diam } G < \infty.$$

Here and below $\|A\|_p = \|A\|_{\mathcal{H}_p \rightarrow \mathcal{H}_p}$ for $A: \mathcal{H}_p \rightarrow \mathcal{H}_p$.

Lemma 2 (see [2]).

The operator $S: \mathcal{H}_p \rightarrow \mathcal{H}_p$ is bounded and the following estimate holds:

$$\|S\|_p \leq \frac{1}{2} \max \left\{ \int_{-1}^1 |W_+(y)| dy; \int_{-1}^1 |W_-(y)| dy \right\}.$$

Lemma 3 (see [2, 3]).

The operator $L^{-1}S: \mathcal{H}_p \rightarrow \mathcal{H}_p$ is completely continuous.

Lemma 4.

Almost everywhere in G let

$$\left[\int_{\Omega \times G} |\xi_{\pm}(s', P') \mu_{\pm}(P, P')|^{r_{\pm}} ds' dP' \right]^{1/r_{\pm}} \leq C_1^{\pm}, \quad (4)$$

$$\left[\int_{\Omega \times G} |\eta_{\pm}(s', P') \mu_{\pm}(P, P')|^{\omega_{\pm}} ds' dP' \right]^{1/\omega_{\pm}} \leq C_2^{\pm} \quad (5)$$

and almost everywhere in $\Omega \times G$ let

$$\left[\int_G |\xi_{\pm}(s, P) \mu_{\pm}(P', P)|^{\sigma_{\pm}} dP' \right]^{1/\sigma_{\pm}} \leq C_3^{\pm},$$

$$\left[\int_G |\eta_{\pm}(s, P) \mu_{\pm}(P', P)|^{\delta_{\pm}} dP' \right]^{1/\delta_{\pm}} \leq C_4^{\pm},$$

where $C_1^{\pm}, \dots, C_4^{\pm}$ are constant numbers, $\sigma_{\pm}, \delta_{\pm} \leq p$, $(p - \sigma_{\pm}) / (p - 1) < r_{\pm}$, $(p - \delta_{\pm}) / (p - 1) < \omega_{\pm}$. Then the operators $R_{\pm}: \mathcal{H}_p \rightarrow L_p(G)$ are completely continuous and the following estimates hold

$$\|R_{\pm}\|_{\mathcal{H}_p \rightarrow L_p(G)} \leq 2^{1/p'} \max \left\{ \frac{(C_1^{\pm})^{1-\sigma_{\pm}/p} (C_3^{\pm})^{\sigma_{\pm}/p} (4\pi V)^{1/p' - 1/r_{\pm} + \sigma_{\pm}/p r_{\pm}}}{(\sigma^+)^{1/p}}; \right. \\ \left. \frac{(C_2^{\pm})^{1-\delta_{\pm}/p} (C_4^{\pm})^{\delta_{\pm}/p} (4\pi V)^{1/p' - 1/\omega_{\pm} + \delta_{\pm}/p \omega_{\pm}}}{(\sigma^-)^{1/p}} \right\},$$

where $p' = p / (p - 1)$, $V = \text{mes } G$.

Lemma 5. Let conditions (4), (5) be satisfied for some $r_{\pm}, \omega_{\pm} \geq 1/p'$. Then the operators $R_{\pm}: \mathcal{H}_p \rightarrow L_{\infty}(G)$ are continuous.

Proof of Lemmas 4, 5. We represent the operators $R_{\pm}: \mathcal{H}_p \rightarrow L_q(G)$, $1 \leq q \leq \infty$ in the form $R_{\pm}\psi = R_{\pm}^{(1)}\psi_+ - R_{\pm}^{(2)}\psi_-$, $\psi = \text{col } \{\psi_+, \psi_-\}$, where

$$R_{\pm}^{(1)}\psi_+ = \int_{\Omega \times G} \xi_{\pm}(s', P') \mu_{\pm}(P, P') \psi_+(s', P') ds' dP',$$

$$R_{\pm}^{(2)}\psi_- = \int_{\Omega \times G} \eta_{\pm}(s', P') \mu_{\pm}(P, P') \psi_-(s', P') ds' dP'.$$

It follows from [4] that when the conditions of Lemma 4 are satisfied the operators $R_{\pm}^{(1)}: \mathcal{H}_p^+ \rightarrow L_p(G)$, $R_{\pm}^{(2)}: \mathcal{H}_p^- \rightarrow L_p(G)$ are completely continuous. But then $R_{\pm}: \mathcal{H}_p \rightarrow L_p(G)$ are also completely continuous. Moreover, the conditions of Lemma 5 ensure the continuity of the operators $R_{\pm}^{(1)}: \mathcal{H}_p^+ \rightarrow L_{\infty}(G)$, $R_{\pm}^{(2)}: \mathcal{H}_p^- \rightarrow L_{\infty}(G)$, and therefore also the continuity of the operators $R_{\pm}: \mathcal{H}_p \rightarrow L_{\infty}(G)$. Estimates for the norms $\|R_{\pm}\|_{\mathcal{H}_p \rightarrow L_p(G)}$ follow from the estimate

$$\|R_{\pm}\psi\|_{L_p(G)} \leq \max \left\{ \frac{\|R_{\pm}^{(1)}\|_{L_p(\Omega \times G) \rightarrow L_p(G)}}{(\sigma^+)^{1/p}}; \frac{\|R_{\pm}^{(2)}\|_{L_p(\Omega \times G) \rightarrow L_p(G)}}{(\sigma^-)^{1/p}} \right\} 2^{1/p'} \|\psi\|_p$$

and estimates for the norms $\|R_{\pm}^{(1,2)}\|_{L_p(\Omega \times G) \rightarrow L_p(G)}$, are given in [4].

Lemma 6.

The superposition operator F acts from $L_q(G)$ into $L_q(G)$ for $1 \leq q \leq \infty$, and is continuous and bounded on every sphere in $L_q(G)$ for $1 < q < \infty$. The operators FR_{\pm} are completely continuous from \mathcal{H}_p^0 into $L_p(G)$ when the conditions of Lemma 4 are satisfied, $1 < p < \infty$.

The proof of the lemma follows from the continuity of the function $\mathcal{F}(x)$ (see subsection 3) and the estimates $|\mathcal{F}(x)| \leq |u_0| + |x|$, $x \in (-\infty, \infty)$ (see [5], p. 312).

2. The stationary problem

1. Statement of the problem. We consider the stationary problem for the integro-differential system

$$\begin{aligned} & (s, \nabla) \psi_+(s, P) + \Sigma^+(s, P) \psi_+(s, P) \\ &= \int_{\Omega} W_+(s, s') \Sigma^+(s', P) \psi_+(s', P) ds' + Q^+(s, P) F_+(\psi_+, \psi_-), \\ & (s, \nabla) \psi_-(s, P) + \Sigma^-(s, P) \psi_-(s, P) \\ &= \int_{\Omega} W_-(s, s') \Sigma^-(s', P) \psi_-(s', P) ds' + Q^-(s, P) F_-(\psi_+, \psi_-) \end{aligned} \quad (6)$$

with the boundary condition

$$\psi_{\pm}(s, P) = \varphi'_{\pm}(s, P), \quad P \in \Gamma_{-s}. \quad (7)$$

In what follows we will consider only those boundary conditions $\psi'(s, P)$ which permit continuation onto W_p , that is, a $\varphi \in W_p$, is found such that $\varphi = \varphi'$ for $P \in \Gamma_-$ (on the subject of continuation see [6]). A generalized L_p -solution of the system (6), (7) is defined as a function $\psi(s, P) \in W_p$, for which

$$\mathcal{L}\psi = S\psi + B'\psi, \quad \psi - \varphi \in D_p.$$

Here

$$B'\psi = \text{col} \left\{ \frac{Q^+(s, P)}{\Sigma^+(s, P)} FR_+\psi, \frac{Q^-(s, P)}{\Sigma^-(s, P)} FR_-\psi \right\}.$$

For the purpose of studying system (6), (7) we introduce the following operator equation in the space \mathcal{H}_p :

$$\psi = L^{-1}S\psi + L^{-1}B\psi + f \equiv A\psi, \quad (8)$$

where $B\psi \equiv B'(\psi + \varphi)$, $f = L^{-1}S\varphi - L^{-1}\mathcal{L}\varphi$. The following lemma holds.

Lemma 7.

Let $\chi \in D_p$ be the solution of (8), then $\psi = \chi + \varphi$ is the L_p -solution of the system (6), (7). Conversely, if ψ is the L_p -solution of (6), (7), then $\chi = \psi - \varphi$ satisfies (8).

Therefore, the question of the existence and uniqueness of the L_p -solution of system (6), (7) reduces to the question of the existence and uniqueness of the fixed point of the operator $A: \mathcal{H}_p \rightarrow \mathcal{H}_p$ in \mathcal{H}_p .

2. Lemma 8.

The operator $\mathcal{H}_p \rightarrow \mathcal{H}_p$ satisfies the Lipschitz condition $\|A\psi_1 - A\psi_2\|_p \leq a_p \|\psi_1 - \psi_2\|_p$, $\psi_{1,2} \in \mathcal{H}_p$, where $a_p = \|L^{-1}S\|_p + \|L^{-1}\|_p k_p$,

$$k_p^p = \left[\Sigma^+ \left(\frac{Q^+}{\sigma^+} \right)^p \|R_+\|_{\mathcal{H}_p \rightarrow L_p(G)}^p + \Sigma^- \left(\frac{Q^-}{\sigma^-} \right)^p \|R_-\|_{\mathcal{H}_p \rightarrow L_p(G)}^p \right] 4\pi.$$

Proof. It follows from the form of the function $\mathcal{F}(x)$ (subsection 3), that $|\mathcal{F}(x_1) - \mathcal{F}(x_2)| \leq |x_1 - x_2|$, $x_{1,2} \in (-\infty, \infty)$. Then

$$\begin{aligned} \|B\psi_1 - B\psi_2\|_p^p &= \left\| \frac{Q^+(s, P)}{\Sigma^+(s, P)} [FR_+(\psi_1 + \varphi) - FR_+(\psi_2 + \varphi)] \right\|_p^+{}^p \\ &+ \left\| \frac{Q^-(s, P)}{\Sigma^-(s, P)} [FR_-(\psi_1 + \varphi) - FR_-(\psi_2 + \varphi)] \right\|_p^-{}^p \\ &\leq 4\pi \Sigma^+ \left(\frac{Q^+}{\sigma^+} \right)^p \|R_+(\psi_1 - \psi_2)\|_{L_p(G)}^p + \\ &+ 4\pi \Sigma^- \left(\frac{Q^-}{\sigma^-} \right)^p \|R_-(\psi_1 - \psi_2)\|_{L_p(G)}^p \leq k_p^p \|\psi_1 - \psi_2\|_p^p \end{aligned}$$

and therefore

$$\begin{aligned} \|A\psi_1 - A\psi_2\|_p &\leq \|L^{-1}S(\psi_1 - \psi_2)\|_p + \|L^{-1}(B\psi_1 - B\psi_2)\|_p \\ &\leq a_p \|\psi_1 - \psi_2\|_p. \end{aligned}$$

Theorem 1.

Let

$$a_p < 1, \quad (9)$$

then Eq. (8) (system (6), (7)) has a unique solution.

The proof follows from the principle of compressible mappings [5]. Condition (9) can be checked numerically by using the estimates given in Lemmas 1-4.

3. We note that when the conditions of Lemma 4 are satisfied the non-linear operator $A: \mathcal{H}_v \rightarrow \mathcal{H}_p$ is completely continuous. We explain the conditions for which Schauder's principle holds for A [5]. Let $\|\psi + \varphi\|_p \leq r$, then

$$\|A\psi + \varphi\|_p \leq \|L^{-1}S\|_p r + \|L^{-1}\|_p \|B\psi\|_p + \|\varphi - L^{-1}\mathcal{L}\varphi\|_p.$$

Since

$$\begin{aligned} \|B\psi\|_p &\leq 4\pi \Sigma^+ \left(\frac{Q^+}{\sigma^+} \right)^p \|R_+\|_{\mathcal{H}_p \rightarrow L_p(G)}^p r^p \\ &+ 4\pi \Sigma^- \left(\frac{Q^-}{\sigma^-} \right)^p \|R_-\|_{\mathcal{H}_p \rightarrow L_p(G)}^p r^p \\ &+ \left\{ |u_0|^p \int_{\Omega \times G} \left[\Sigma^+(s, P) \left| \frac{Q^+(s, P)}{\Sigma^+(s, P)} \right|^p \right. \right. \\ &\left. \left. + \Sigma^-(s, P) \left| \frac{Q^-(s, P)}{\Sigma^-(s, P)} \right|^p \right] ds dP \right\} = k_p r^p + \{c_p\}, \end{aligned}$$

therefore $\|A\psi + \varphi\|_p \leq \|L^{-1}S\|_p r + \|\varphi - L^{-1}\mathcal{L}\varphi\|_p + \|L^{-1}\|_p \{k_p r^p + c_p\}^{1/p} = b_p$. Therefore, if $b_p \leq r$, then the operator A maps the sphere $\|\psi + \varphi\|_p \leq r$ into itself and Theorem 2 holds.

Theorem 2.

Let an $r > 0$, be found such that

$$b_p \leq r, \quad (10)$$

then Eq. (8) has at least one solution in the sphere $\|\psi + \varphi\|_p \leq r$.

However, it is easy to show that conditions (9) and (10) are equivalent: (10) implies (9) and conversely, when (9) is satisfied an $r_p > 0$, can be found such that (10) is satisfied for all $r \geq r_p$. Nevertheless Theorem 2 permits us to obtain for the L_p -solution of (6), (7) an *a priori* estimate in terms of the data of the problem, namely: for $a_p < 1$

$$\|\psi\|_p \leq r_p,$$

where

$$\begin{aligned} r_p &\leq \inf_{0 < \varepsilon < (1 - \|L^{-1}S\|_p) / \|L^{-1}\|_p - k_p} \max \left\{ \frac{\|\varphi - L^{-1}\mathcal{L}\varphi\|_p}{\varepsilon}; \right. \\ &\left. c_p \left[\left(\frac{1 - \|L^{-1}S\|_p}{\|L^{-1}\|_p} - \varepsilon \right)^p - k_p \right]^{-1/p} \right\}. \end{aligned}$$

4. For practical estimates it is interesting to know beforehand, in terms of the data of the problem, the behaviour relative to each other of the components of the solution $\psi = \text{col} \{\psi_+, \psi_-\}$ and also the conditions ensuring the existence of a solution non-negative almost everywhere in $\Omega \times G$.

Let $K \subset \mathcal{H}_p$ be a cone of the form

$$K = \{x = \text{col} \{x_+, x_-\} \in \mathcal{H}_p: x_+(s, P) \geq x_-(s, P) \geq 0 \text{ almost everywhere in } \Omega \times G\}.$$

We introduce the linear operator $R: \mathcal{H}_p \rightarrow \mathcal{H}_p$ as follows:

$$Rx = \text{col} \left\{ \frac{Q^+(s, P)}{\Sigma^+(s, P)} R_+ x; \frac{Q^-(s, P)}{\Sigma^-(s, P)} R_- x \right\}, \quad x \in \mathcal{H}_p.$$

Lemma 9.

Let the conditions of Lemma 4 hold. Then the operator $R: \mathcal{H}_p \rightarrow \mathcal{H}_p$ is completely continuous and $\|R\|_p \leq k_p$.

We suppose that almost everywhere in the corresponding domains of definition the following conditions are satisfied:

$$\begin{aligned} \xi_+(s, P) &\geq \xi_-(s, P) \geq \eta_-(s, P) \geq \eta_+(s, P) \geq 0, \\ \mu_+(P, P') &\geq \mu_-(P, P') \geq 0, \\ W_+(s \cdot s') \Sigma^+(s, P - \xi s) \exp \left[- \int_0^\xi \Sigma^+(s, P - \xi' s) d\xi' \right] \\ &\geq W_-(s \cdot s') \Sigma^-(s, P - \xi s) \exp \left[- \int_0^\xi \Sigma^-(s, P - \xi' s) d\xi' \right] \geq 0, \\ Q^+(s, P - \xi s) \exp \left[- \int_0^\xi \Sigma^+(s, P - \xi' s) d\xi' \right] \\ &\geq Q^-(s, P - \xi s) \exp \left[- \int_0^\xi \Sigma^-(s, P - \xi' s) d\xi' \right] \geq 0, \quad 0 \leq \xi \leq d. \end{aligned} \quad (11)$$

Lemma 10.

Let K_L be the cone of non-negative functions in $L_p(G)$. Then when (11) is satisfied, $x \in K$ implies $R_\pm x \in K_L$.

Lemma 11.

Let conditions (11) hold and $\varphi, f \in K$. Then the operator A is positive on K .

The proof follows from the monotonicity of the function $\mathcal{F}(x)$ and the representations for the operators $L_\pm^{-1} S_\pm$, L_\pm^{-1} , given in [2].

Lemma 12.

Let $u_0 \geq 0$ (see section 1, subsection 3) and the conditions of Lemma 11 be satisfied. Then the operator A has a strong asymptotic derivative $A'(\infty)$ with respect to the cone K [5] and $A'(\infty) = L^{-1} S + L^{-1} R$.

Proof. Let $x \in K$, then

$$\|Ax - A'(\infty)x\|_p \leq \|f\|_p + \|L^{-1}\|_p \|Bx - Rx\|_p. \quad (12)$$

Since $R_{\pm}(x+\varphi) \in K_L$ for $x, \varphi \in K$ (Lemma 10)) and consequently the functions $R_{\pm}(P) = R_{\pm}(x+\varphi)$ are non-negative almost everywhere in G , then $\mathcal{F}(R_{\pm}(P)) = R_{\pm}(P) + u_0$ almost everywhere in G . Therefore

$$\|Bx - Rx\|_p^p = \left[\left\| \frac{Q^+(s, P)}{\Sigma^+(s, P)} (R_+ \varphi + u_0) \right\|_p^+ \right]^p + \left[\left\| \frac{Q^-(s, P)}{\Sigma^-(s, P)} (R_- \varphi + u_0) \right\|_p^- \right]^p.$$

Therefore (12) implies that

$$\lim_{R \rightarrow \infty} \sup_{\|x\|_p \geq R, x \in K} \frac{\|Ax - A'(\infty)x\|_p}{\|x\|_p} = 0,$$

which is what is required to prove.

Theorem 3 (see [5], p. 419).

Let A be a completely continuous positive operator, $A'(\infty)$ its strong asymptotic derivative. If the spectral radius of the operator $A'(\infty)$ is less than 1, then the operator A has in K at least one fixed point.

Lemmas 9–12 and Theorem 3, which we have proved, enable us to make the following statement.

Theorem 4.

Let the conditions of Lemmas 4, 12 and

$$\mathcal{R} = \lim_{n \rightarrow \infty} [\|(L^{-1}S + L^{-1}R)^n\|_p]^{1/n} < 1. \quad (13)$$

be satisfied. Then Eq. (8) has at least one solution in K .

Corollary 1. We note (see Lemma 9), that $\mathcal{R} \leq \|L^{-1}S + L^{-1}R\|_p \leq \|L^{-1}S\|_p + \|L^{-1}\|_p k_p = a_p$. Consequently, (13) holds if (9) is satisfied. Therefore subject to conditions (9) and (11) the unique L_p -solution of the system (6), (7) $\psi = \text{col} \{\psi_+, \psi_-\}$ possesses the property $\psi_+(s, P) \geq \psi_-(s, P) \geq 0$ almost everywhere in $\Omega \times G$.

Corollary 2. If in (11) the signs \geq in the right sides of the inequalities are replaced by the sign $=$, then if $a_p < 1$ for a unique L_p -solution of system (6), (7) we have $\psi_+(s, P) = \psi_-(s, P) \geq 0$ almost everywhere in $\Omega \times G$.

3. The non-stationary problem

1. Statement of the problem. In this subsection we study the non-stationary problem (1)–(3). We assume that the boundary conditions $\varphi'(s, P, t)$ permit continuation onto W_p , that is, for all $t \in [0, T]$, a $\varphi(t) \in W_p$ can be found such that $\varphi(t) = \varphi'(t)$ for $P \in \Gamma_{-s}$, $t \in [0, T]$.

We define a generalized L_p -solution of problem (1)–(3) as a mapping $\psi(t) : [0, T] \rightarrow W_p$, possessing in \mathcal{H}_p for all $t \in [0, T]$ a continuous strong derivative $d\psi/dt$, for which

$$d\psi/dt + \mathcal{L}_1 \psi = C_1 \psi, \quad \psi(t) - \varphi(t) \in D_p, \quad \psi(0) = \psi^{(0)} \in W_p$$

is satisfied for all $t \in [0, T]$. Here

$$\begin{aligned} \mathcal{L}_1 &= \text{diag} \{ \Sigma^+(s, P) \mathcal{L}_+, \Sigma^-(s, P) \mathcal{L}_- \}, \\ C_1 \psi &= \text{col} \{ \Sigma^+(s, P) S_+ \psi_+ + Q^+(s, P) FR_+ \psi, \\ &\quad \Sigma^-(s, P) S_- \psi_- + Q^-(s, P) FR_- \psi \}. \end{aligned}$$

We suppose that $\varphi(t)$ has in \mathcal{H}_p for all $t \in [0, T]$ a continuous strong derivative $d\varphi/dt$. For the purpose of studying system (1)–(3) we consider in \mathcal{H}_p the following non-stationary equation:

$$d\psi/dt + L_1 \psi = C(t, \psi), \quad \psi(0) = \psi^{(0)} \in D_p, \quad (14)$$

where

$$\begin{aligned} L_1 &= \text{diag} \{ \Sigma^+(s, P) L_+, \Sigma^-(s, P) L_- \}, \\ C(t, \psi) &= \text{col} \{ \Sigma^+(s, P) S_+ \psi_+ + Q^+(s, P) FR_+ (\psi + \varphi(t)) \\ &\quad + \Sigma^+(s, P) (S_+ - \mathcal{L}_+) \varphi_+(t) - d\varphi_+/dt, \\ &\quad \Sigma^-(s, P) S_- \psi_- + Q^-(s, P) FR_- (\psi + \varphi(t)) \\ &\quad + \Sigma^-(s, P) (S_- - \mathcal{L}_-) \varphi_-(t) - d\varphi_-/dt \}. \end{aligned}$$

The solution of (14) is understood in the ordinary sense [7].

Lemma 13.

Let $\chi(t)$ be a solution of (14) with $\psi^{(0)} = \psi^{(0)} - \varphi(0)$, then $\psi(t) = \chi(t) + \varphi(t)$ is an L_p -solution of the system (1)–(3). Conversely, if $\psi(t)$ is an L_p -solution of (1)–(3), then $\chi(t) = \psi(t) - \varphi(t)$ is a solution of (14) with $\psi^{(0)} = \psi^{(0)} - \varphi(0)$.

Therefore, the question of the L_p -solvability of system (1)–(3) has been reduced to the question of the existence and uniqueness of the solution of (14).

Lemma 14 (see [8]).

The operator $-L_1$ generates a C_0 -semigroup $U(t)$ in \mathcal{H}_p .

2. In this subsection it is assumed that the function $\mathcal{F}(x)$ (section 1, subsection 3) in an ϵ -neighbourhood of the point $-u_0$ is compressed in such a way that the compressed function $\mathcal{F}(x) \in C^2$ satisfies the conditions

$$\begin{aligned} \widetilde{\mathcal{F}}(x) &\equiv \mathcal{F}(x) \text{ if } |x + u_0| > \epsilon \text{ and } x < -u_0 - \epsilon, \\ \max_{-\infty < x < \infty} |\widetilde{\mathcal{F}}''(x)| &= k < \infty. \end{aligned}$$

Lemma 15.

We suppose that $\varphi(t)$ has in W_p for $t \in [0, T]$ a strong derivative $d\varphi/dt$ continuous with respect to t , and that $d\varphi/dt$ has in \mathcal{H}_p a continuous strong derivative $d^2\varphi/dt^2$ for $t \in [0, T]$. Also let the conditions of Lemma 5 be satisfied. Then derivatives $\tilde{C}_t'(t, x)$, $\tilde{C}_x'(t, x)$, continuous with respect to the ensemble of variables, of the operator $\tilde{C}(t, x)$ exist:

$$\begin{aligned} \tilde{C}_t'(t, x) = & \text{col} \left\{ Q^+(s, P) F' R_+(x + \varphi(t)) R_+ \left(\frac{d\varphi}{dt} \right) \right. \\ & + \Sigma^+(s, P) [S_+ - \mathcal{L}_+] \left(\frac{d\varphi_+}{dt} \right) \\ & - \frac{d^2\varphi_+}{dt^2}, Q^-(s, P) F' R_-(x + \varphi(t)) R_- \left(\frac{d\varphi}{dt} \right) \\ & \left. + \Sigma^-(s, P) [S_- - \mathcal{L}_-] \left(\frac{d\varphi_-}{dt} \right) - \frac{d^2\varphi_-}{dt^2} \right\}, \end{aligned} \quad (15)$$

$$\begin{aligned} & \tilde{C}_x'(t, x) \\ = & \left\| \begin{array}{cc} \Sigma^+(s, P) S_+ + Q^+(s, P) F' R_+(x + \varphi(t)) R_+ & Q^+(s, P) F' R_+(x + \varphi(t)) R_+ \\ Q^-(s, P) F' R_-(x + \varphi(t)) R_- & \Sigma^-(s, P) S_- + Q^-(s, P) F' R_-(x + \varphi(t)) R_- \end{array} \right\|. \end{aligned} \quad (16)$$

In addition, $\tilde{C}_t'(t, x)$, $\tilde{C}_x'(t, x)$ satisfy a Lipschitz condition on x ($\tilde{C}_t'(t, x)$ in the norm $\|\cdot\|_p$, and $\tilde{C}_x'(t, x)$ in the norm $\|\cdot\|_{\mathcal{H}_p \rightarrow \mathcal{H}_p}$).

The tilde over the operator symbol denotes the substitution $\mathcal{F}(x) \rightarrow \tilde{\mathcal{F}}(x)$; $F'\psi = \tilde{\mathcal{F}}'(\psi(P))$, $\psi(P) \in L_q(G)$.

Proof. Let $x \in \mathcal{H}_p$. We write $\Phi(t) = (S_1 - \mathcal{L}_1)\varphi(t) - d\varphi/dt$, $\Phi'(t) = (S_1 - \mathcal{L}_1)(d\varphi/dt) - d^2\varphi/dt^2$, $S_1 = \text{diag} \{ \Sigma^+(s, P) S_+, \Sigma^-(s, P) S_- \}$. We consider the norm of the difference

$$\begin{aligned} & \left\| \frac{\tilde{C}(t + \Delta t, x) - \tilde{C}(t, x)}{\Delta t} - \tilde{C}_t'(t, x) \right\|_p \\ & \leq \left\| \frac{\Phi(t + \Delta t) - \Phi(t)}{\Delta t} - \Phi'(t) \right\|_p \\ & + \left\| Q^+(s, P) \left[\frac{F R_+(x + \varphi(t + \Delta t)) - F R_+(x + \varphi(t))}{\Delta t} \right. \right. \\ & \quad \left. \left. - F' R_+(x + \varphi(t)) R_+ \left(\frac{d\varphi}{dt} \right) \right] \right\|_p^+ \\ & + \left\| Q^-(s, P) \left[\frac{F R_-(x + \varphi(t + \Delta t)) - F R_-(x + \varphi(t))}{\Delta t} \right. \right. \\ & \quad \left. \left. - F' R_-(x + \varphi(t)) R_- \left(\frac{d\varphi}{dt} \right) \right] \right\|_p^- = \|I_1\|_p + \|I_2\|_p^+ + \|I_3\|_p^-. \end{aligned} \quad (17)$$

Because of the conditions of the present lemma and the continuity of the operators S_1, \mathcal{L}_1 on W_p (Lemma 1), $\|I_1\|_p \rightarrow 0$ as $\Delta t \rightarrow 0$. Since $\tilde{\mathcal{F}}(x) \in C^2$, then for every Δt we can find a function $\theta_{\Delta t}(P)$, $0 \leq \theta_{\Delta t}(P) \leq 1$, measurable in G , such that $\tilde{\mathcal{F}}(R_+(x + \varphi(t + \Delta t))) - \tilde{\mathcal{F}}(R_+(x + \varphi(t))) = \tilde{\mathcal{F}}'(R_+(x + \varphi(t)) + \theta_{\Delta t}(P)[R_+(\varphi(t + \Delta t) - \varphi(t))]) R_+(\varphi(t + \Delta t) - \varphi(t))$ almost everywhere in G .

Because of the boundedness of the operator $R_+ : \mathcal{H}_p \rightarrow L_p(G)$, the strong continuity in t of the function $\varphi(t)$ and the continuity of $\tilde{\mathcal{F}}'(x)$, we have as $\Delta t \rightarrow 0$

$$\begin{aligned} A_1(\Delta t) &\equiv \tilde{\mathcal{F}}'(R_+(x + \varphi(t)) + \theta_{\Delta t}(P)[R_+(\varphi(t + \Delta t) - \varphi(t))]) \\ &\rightarrow \tilde{\mathcal{F}}'(R_+(x + \varphi(t))) \equiv A_2 \end{aligned} \quad (18)$$

with respect to the measure on G .

Adding to and subtracting from I_2 an expression of the form $A_1(\Delta t)R_+(d\varphi/dt)$, it is easy to obtain the following estimate:

$$\begin{aligned} \|I_2\|_p^+ &\leq Q^+ \left\{ 4\pi C \|R_+\|_{\mathcal{H}_p \rightarrow L_p(G)} \left\| \frac{\varphi(t + \Delta t) - \varphi(t)}{\Delta t} - \frac{d\varphi}{dt} \right\|_p \right. \\ &\quad \left. + \left\| [A_1(\Delta t) - A_2]R_+ \left(\frac{d\varphi}{dt} \right) \right\|_p^+ \right\}. \end{aligned}$$

From the estimate $|A_1(\Delta t) - A_2| < C$, where C is a constant number, and (18) we have by Lebesgue's theorem, $\|I_2\|_p^+ \rightarrow 0$ as $\Delta t \rightarrow 0$. Similarly, $\|I_3\|_p^- \rightarrow 0$ as $\Delta t \rightarrow 0$. Therefore, (17) implies the validity of (15). We now show that (16) holds. Let $x, h \in \mathcal{H}_p$. We have

$$\begin{aligned} &\|\mathcal{C}(t, x+h) - \mathcal{C}(t, x) - \mathcal{C}'_x(t, x)h\|_p \\ &\leq \|Q^+(s, P)[\tilde{\mathcal{F}}(R_+(\varphi(t) + x+h)) - \tilde{\mathcal{F}}(R_+(\varphi(t) + x)) \\ &\quad - \tilde{\mathcal{F}}'(R_+(\varphi(t) + x))R_+h]\|_p^+ + \|Q^-(s, P)[\tilde{\mathcal{F}}(R_-(\varphi(t) + x+h)) \\ &\quad - \tilde{\mathcal{F}}(R_-(\varphi(t) + x)) - \tilde{\mathcal{F}}'(R_-(\varphi(t) + x))R_-h]\|_p^- \\ &= \|I_4\|_p^+ + \|I_5\|_p^-. \end{aligned} \quad (19)$$

Since $\tilde{\mathcal{F}}(x) \in C^2$, then for every $h \in \mathcal{H}_p$ a function $\theta_h(P)$, $0 \leq \theta_h(P) \leq 1$, measurable on G can be found such that

$$\begin{aligned} &\tilde{\mathcal{F}}(R_+(\varphi(t) + x+h)) - \tilde{\mathcal{F}}(R_+(\varphi(t) + x)) \\ &= \tilde{\mathcal{F}}'(R_+(\varphi(t) + x) + \theta_h(P)R_+h)R_+h. \end{aligned} \quad (20)$$

By the boundedness of $R_+ : \mathcal{H}_p \rightarrow L_p(G)$ and the continuity of $\tilde{\mathcal{F}}'(x)$ as $\|h\|_p \rightarrow 0$, we have

$$A_3(h) \equiv \tilde{\mathcal{F}}'(R_+(\varphi(t) + x) + \theta_h(P)R_+h) \rightarrow \tilde{\mathcal{F}}'(R_+(\varphi(t) + x)) \equiv A_4, \quad (21)$$

with respect to the measure on G . Substituting (20) into I_4 , we obtain the estimate

$$\begin{aligned} \|I_4\|_p^+ &\leq Q^+ \|A_3(h) - A_4\|_p^+ \|R_+h\|_{L_\infty(G)} \\ &\leq Q^+ \|R_+\|_{\mathcal{H}_p \rightarrow L_\infty(G)} \|A_3(h) - A_4\|_p^+ \|h\|_p. \end{aligned} \quad (22)$$

From (21), (22) and the estimate $|A_3(h) - A_4| < C$ we obtain by Lebesgue's theorem that $\|I_4\|_p^+ / \|h\|_p \rightarrow 0$ as $\|h\|_p \rightarrow 0$. Similarly, $\|I_5\|_p^- / \|h\|_p \rightarrow 0$ as $\|h\|_p \rightarrow 0$. By (19), this implies the validity of (16). To prove the last assertion of the lemma we obtain the necessary estimates. Let $x, y, h \in \mathcal{H}_p$, then

$$\begin{aligned} & \|\tilde{C}_t'(t, x) - \tilde{C}_t'(t, y)\|_p \\ & \leq Q^+ \left\| R_+ \left(\frac{d\varphi}{dt} \right) \right\|_{L_\infty(G)} \|FR_+(\varphi(t) + x) - FR_+(\varphi(t) + y)\|_{p^+} \\ & + Q^- \left\| R_- \left(\frac{d\varphi}{dt} \right) \right\|_{L_\infty(G)} \|FR_-(\varphi(t) + x) - FR_-(\varphi(t) + y)\|_{p^-} \\ & \leq (4\pi)^{1/p} k \left[Q^+(\Sigma^+)^{1/p} \left\| R_+ \left(\frac{d\varphi}{dt} \right) \right\|_{L_\infty(G)} \|R_+\|_{\mathcal{H}_p \rightarrow L_p(G)} \right. \\ & \left. + Q^-(\Sigma^-)^{1/p} \left\| R_- \left(\frac{d\varphi}{dt} \right) \right\|_{L_\infty(G)} \right. \\ & \left. \times \left\| R_- \left(\frac{d\varphi}{dt} \right) \right\|_{L_\infty(G)} \|R_-\|_{\mathcal{H}_p \rightarrow L_p(G)} \right] \|x - y\|_p \leq M_p \|x - y\|_p, \\ & M_p = \text{const.} \end{aligned}$$

Similarly,

$$\begin{aligned} & \|[\tilde{C}_x'(t, x) - \tilde{C}_x'(t, y)]h\|_p \\ & \leq (4\pi)^{1/p} k [Q^+(\Sigma^+)^{1/p} \|R_+\|_{\mathcal{H}_p \rightarrow L_\infty(G)} \|R_+\|_{\mathcal{H}_p \rightarrow L_p(G)} \\ & + Q^-(\Sigma^-)^{1/p} \|R_-\|_{\mathcal{H}_p \rightarrow L_\infty(G)} \|R_-\|_{\mathcal{H}_p \rightarrow L_p(G)}] \|x - y\|_p \|h\|_p. \end{aligned}$$

Finally, we note that the continuity of $\tilde{C}_t'(t, x)$, $\tilde{C}_x'(t, x)$ with respect to the ensemble of variables follows from the continuity of $\tilde{\mathcal{F}}'(x)$ and the conditions on $\varphi(t)$.

Theorem 5 (see [7]).

Let the operator $-L_1$ generate a C_0 -semigroup in \mathcal{H}_p . Let $\tilde{A}(t, x)$ have on $[0, T] \times \mathcal{H}_p$ partial derivatives $\tilde{C}_t'(t, x)$, $\tilde{C}_x'(t, x)$ (in the Frechét sense), continuous with respect to the ensemble of variables, satisfying a Lipschitz condition on x . Then there exists a solution of (14) defined on some segment $[0, T_1] \subset [0, T]$.

Theorem 6.

Let the conditions of Lemma 7 be satisfied. The (14) has (for $\tilde{\mathcal{F}}(x)$) a unique solution defined on some segment $[0, T_1] \subset [0, T]$.

3. Let $\mathcal{F}(x)$ be the function introduced in section 1, subsection 3. Let $\varphi'(s, P, t) \equiv \varphi'(s, P)$, $t \in [0, T]$. Then in $\tilde{L}_2 = L_2(\Omega \times G) \times L_2(\Omega \times G)$ we can establish Theorem 8 of the existence and uniqueness of the solution of (14), which is a corollary of Pao's result (Theorem 7) and Lemmas 16, 17.

Theorem 7 (see [9]).

Let $-L_1$ be dissipative in \tilde{L}_2 with constant β and the domain of values of the operator

$\alpha I + L_1$, $\alpha > \beta$, be identical with \tilde{L}_2 . If $C(t, \psi) \equiv C(\psi)$ satisfies the Lipschitz condition

$$\|C(\psi) - C(\varphi)\|_{\tilde{L}_2} \leq k_1 \|\psi - \varphi\|_{\tilde{L}_2}, \quad \psi, \varphi \in \tilde{L}_2, \quad k_1 = \text{const},$$

then (14) has a unique solution. If in addition $\beta > 0$ and $k_1 < \beta$, then every equilibrium solution (see [9]), if it exists, is exponentially asymptotically stable.

Lemma 16.

The operator $-L_1$ is strictly dissipative in \tilde{L}_2 .

The proof follows from the estimate $(L_1\varphi, \varphi) = (L_1\varphi_+, \varphi_+) + (L_1\varphi_-, \varphi_-) = (L_+\varphi_+, \varphi_+) + (L_-\varphi_-, \varphi_-) \geq (\varphi_+, \varphi_+) + (\varphi_-, \varphi_-) = \|\varphi\|_2^2 \geq a^2 \|\varphi\|_{\tilde{L}_2}^2$, which follows from the result (see [2]) $(L_{\pm}\varphi_{\pm}, \varphi_{\pm})_{\pm} \geq (\varphi_{\pm}, \varphi_{\pm})_{\pm}$. Here

$$a^2 = \min\{\sigma^+, \sigma^-\}, \quad (x, y)_{\pm} = \int_{\Omega \times G} \Sigma^{\pm}(s, P) x(s, P) y(s, P) ds dP.$$

Lemma 17.

The operator $C: \mathcal{H}_p \rightarrow \mathcal{H}_p$ satisfies a Lipschitz condition with constant

$$k_1 = \frac{b^3}{a} \|S\|_2 + b[(Q^+)^2 \|R_+\|_{\mathcal{H}_2 \rightarrow L_2(G)}^2 + (Q^-)^2 \|R_-\|_{\mathcal{H}_2 \rightarrow L_2(G)}^2]^{1/2},$$

where $b^2 = \max\{\Sigma^+, \Sigma^-\}$.

The proof of the lemma is similar to that of Lemma 8, if we take into account the fact that $a \|\varphi\|_{\tilde{L}_2} \leq \|\varphi\|_2 \leq b \|\varphi\|_{\tilde{L}_2}$.

Theorem 8.

Let $\varphi'(s, P, t) \equiv \varphi'(s, P)$, $t \in [0, T]$. Then (14) has in \tilde{L}_2 a unique solution in any interval $[0, T]$. Moreover, if $k_1 < a^2$, then every equilibrium solution [9], if it exists, is exponentially asymptotically stable.

Translated by J. Berry.

REFERENCES

1. VARSHAVSKII, V. I. On the mathematical theory of neuron meshes, In: *Applications of mathematical methods in biology* (Primenenie matem. metodov v biologii), II, 60-66, Izf-vo LGU, Leningrad, 1963.
2. VLADIMIROV, V. S. Mathematical problems of the single-velocity theory of particle transfer. *Tr. Matem. on-ta Akad. Nauk USSR*, 61, 1961.
3. GERMOGENOVA, T. A. Generalized solutions of boundary value problems for transfer equations. *Zh. vychisl. Mat. mat. Fiz.*, 9, 3, 605-625, 1969.
4. KANTOROVICH, L. V. and AKILOV, G. P. *Functional analysis in normed spaces* (Funktsional'nyi analiz v normirovannykh prostranstvakh). Fizmatgiz, Moscow, 1959.
5. *Functional analysis* (Funktsional'nyi analiz). SMB. "Nauka", Moscow, 1972.
6. SLOBODETSKII, L. I. The generalized spaces of S. L. Sobolev and their applications to boundary value problems for partial differential equations. *Uch. zap. fiz.-matem. fak-ta LGPI im. A. I. Gertsena*, 54-112, 1958.
7. KRASNOSEL'SKII, M. A., KREIN, S. G. and SOBOLEVSKII, P. E. On differential equations with unbounded operators in Banach spaces. *Dokl. Akad. Nauk SSSR*, 111, 1, 19-22, 1956.

8. LUZNETSOV, Yu. A. and MOROZOV, S. F. Mathematical problems of reactor kinetics. *Dokl. Akad. Nauk SSSR*, **218**, 543-546, 1974.
9. PAO, S. V. On nonlinear neutron transport equations. *J. Math. Anal. Appl.*, **42**, 578-593, 1973.

THE SPATIAL KINETICS OF A PULSED HEAT-CAPACITY REACTOR*

A. D. KLIMOV, L. G. STRAKHOVSKAYA, R. P. FEDORENKO and
I. L. CHIKHLADZE

Moscow

(Received 24 April 1975; revised 18 September 1975)

A METHOD of integrating the three-dimensional kinetic equations describing the evolution of the neutron and temperature fields during a neutron burst in a pulsed heat-capacity reactor is presented.

1. Statement of the problem

Until recently transfer processes in nuclear reactors in the majority of cases were described in the point model approximation (see [1]). However, this approximation was often found to be inapplicable.

A number of papers exist devoted to the development and extension of various methods of studying three-dimensional kinetics. The numerical methods that are most promising in their possibilities have been intensively applied in connection with the emergence of high-powered computers [2].

In this paper we present a method integrating the non-stationary diffusion equations describing the kinetics of the neutron field of pulsed heat-capacity reactor [3]. The method considered can be generalized to solve a wider class of applied problems of reactor physics. In particular, it makes it possible to investigate the kinetics of the neutron field in the control of criticality or of the neutron flux at a fixed point, to study depletion processes in systems allowing for local properties at each point of the active zone etc.

To be specific we consider a neutron pulsed reactor or the IGR type in (r, z) -geometry. Before ignition the reactor is in the critical state with minimum controlled level of the neutron flux Φ_0 and temperature T_0 . To produce the pulse a reactivity jump is applied to the reactor. The energy emitted in the nuclear reactor is stored in the graphite of the active zone as heat, and as the stack heats up the burst is quenched because of the negative temperature effect.

In the initial stage of development of the pulse the steady period of the start up is much less than the life-time of the sources of the delayed neutrons, so their effect can be neglected. Below their effect appears in the fact that the decay of the neutron flux proceeds more slowly.

For simplicity we consider a model in which the evolution of the neutron field is described in the two-group diffusion approximation:

$$\frac{1}{v_1} \frac{\partial \Phi_1}{\partial t} = \text{div } D_1 \nabla \Phi_1 - \Sigma_{a1} \Phi_1 + (1-\beta) v \Sigma_f \Phi_2 + \sum_{i=1}^6 \lambda_i C_i, \quad (1.1)$$

*Zh. vychisl. Mat. mat. Fiz., **17**, 1, 162-174, 1977.

$$\frac{1}{v_2} \frac{\partial \Phi_2}{\partial t} = \text{div } D_2 \nabla \Phi_2 - \Sigma_{a2} \Phi_2 + \Sigma_{a1}' \Phi_1, \quad (1.2)$$

$$\frac{\partial C_i}{\partial t} = -\lambda_i C_i + \beta_i \nu \Sigma_f \Phi_2, \quad i=1, 2, \dots, 6, \quad (1.3)$$

$$\frac{\partial T}{\partial t} = \frac{a \Sigma_f}{\gamma c(T)} \Phi_2, \quad (1.4)$$

where Φ_1 is the flux of fast neutrons, $\text{cm}^{-2}\text{sec}^{-1}$, Φ_2 is the flux of thermal neutrons, $\text{cm}^{-2}\text{sec}^{-1}$, C_i is the density of the sources of delayed neutrons of the i -th type, cm^{-3} , T is the temperature of the medium, $^\circ\text{K}$, D_1, D_2 are the diffusion coefficients for fast and thermal neutrons, cm , Σ_{a1}, Σ_{a2} is the absorption section for neutrons of the corresponding energy groups, cm^{-2} , Σ_f is the fission cross-section, cm^{-2} , ν is the number of secondary neutrons, Σ_{a1}' is the "withdrawal" cross-section of the fast neutrons, β_i is the fraction of delayed neutrons of the i -th type, β is the effective fraction of the delayed neutrons, λ_i is the decay constant of delayed neutrons of the i -th type, sec^{-1} , $c(T)$ is the specific heat of graphite, cal/degree , γ is the density of the graphite, g/cm^3 , $a = 7.258 \times 10^{-12} \text{ cal/fission}$, and ν_1, ν_2 are the effective velocities of the corresponding energy groups of neutrons. [For thermal systems of the IGR type the approximation considered is fairly complete; the generalization to the case of the multigroup diffusion approximation is a formal standard procedure.]

The coefficients $D_i, \Sigma_{a1}, \Sigma_f, \nu_2, c(T)$ are functions of the coordinates and temperature $T(\bar{r}, t)$.

Since after the characteristic time of development of the neutron pulse $\Delta t \leq 1 \text{ sec}$ the temperature adjusts itself because of the thermal conductivity at distances $l \sim (\Delta t a)^{1/2} \ll R_{a.z.}$ (where a is the thermal conductivity of the active zone $\text{cm}^2\text{sec}^{-1}$, and $R_{a.z.}$ is the radius of the active zone), then in Eq. (1.4) we neglect the effect of thermal conductivity. There is no difficulty in allowing for this effect in (1.4) within the limits of the method considered.

The functions $\Phi_j(\bar{r}, t), C_i(\bar{r}, t), T(\bar{r}, t)$ satisfy the initial and boundary conditions

$$\Phi_j(\bar{r}, t) |_{t=0} = \Phi_{0j} = \text{const}, \quad j=1, 2, \quad (1.5)$$

$$C_i(\bar{r}, t) |_{t=0} = 0, \quad i=1, 2, \dots, 6, \quad (1.6)$$

$$T(\bar{r}, t) |_{t=0} = T_0, \quad (1.7)$$

$$\frac{\partial}{\partial r} \Phi_j(\bar{r}, t) |_{r=0} = 0, \quad j=1, 2, \quad (1.8)$$

$$\Phi_j(\bar{r}, t) |_{r=0} = 0, \quad j=1, 2, \quad (1.9)$$

where Γ is the extrapolation boundary of the reactor.

Equation (1.3) is eliminated by integration and substitution of the corresponding terms in Eq. (1.1):

$$\sum_{i=1}^6 \lambda_i C_i(\bar{r}, t) = \nu \sum_{i=1}^6 \beta_i \lambda_i \exp(-\lambda_i t) \int_0^t \exp(\lambda_i t') \Sigma_f(\bar{r}, t') \Phi_2(\bar{r}, t') dt'. \quad (1.10)$$

The choice of the method for the numerical solution of the problem is essentially determined by the characteristic features of the process of development of the neutron pulse described by Eqs. (1.1)–(1.9). These features are as follows.

1. Three characteristic time scales exist, different from each other: $\tau_1 \ll \tau_2 \ll \tau$, where τ_1 is the characteristic time for Eq. (1.1), τ_2 is that for Eq. (1.2), and τ is the characteristic time of the process considered. The relation $\tau_1 \ll \tau_2$ is due to the fact that $\nu_1 \gg \nu_2$ ($\nu_1 \approx 10^3 \nu_2$) and after a time $\sim \tau_1$ the function Φ_2 changes negligibly, but Φ_1 becomes "steady-state" as the solution of the equation

$$L_1 \Phi_1 + A_{11} \Phi_1 + A_{12} \Phi_2 \approx 0$$

with a given value of Φ_2 . The relation $\tau_2 \ll \tau$ is connected with the value of the coefficient in (1.4) and with the nature of the dependence of the coefficients of Eqs. (1.1) and (1.2) on the temperature $T(r, z, t)$. Times, small from the point of view of the variation of $T(r, z, t)$, are in this problem large from the point of view of the variation of Φ_2 (and all the more of Φ_1).

2. Discontinuous coefficients. The whole domain of the calculation is subdivided into a comparatively large number of zones, in each of which its own values of the coefficients of the system are specified, the disparity in these values being fairly large. Therefore operators of the type

$$\frac{1}{r} \frac{\partial}{\partial r} r D \frac{\partial}{\partial r}, \quad \frac{\partial}{\partial z} D \frac{\partial}{\partial z}$$

(and their difference approximations) are non-commutative, and this causes certain computational difficulties.

3. At the initial instant the distribution $T(r, z, 0)$, $\Phi_1(r, z, 0)$, $\Phi_2(r, z, 0)$ of the neutron background is specified.

By varying the absorption properties of the medium (withdrawal of the control rods from the active zone) the reactivity varies and the system becomes subcritical. This leads to an exponential increase in the functions Φ_1 and Φ_2 with a period proportional to the subcriticality in the initial state.

On attaining some level of Φ_1 and Φ_2 an increase in temperature by Eq. (1.4) begins. The change in $T(r, z, t)$ leads to a change in the coefficients in Eqs. (1.1), (1.2), the rate of increase of Φ_1 , Φ_2 slows down, and the system gradually passes from the subcritical state into the critical state, after which the increase in Φ_1 , Φ_2 , is replaced by a fall to a value at which the increase of temperature practically ceases: the system passes into the subcritical state. Therefore, after a time of the order of ~ 1 sec the functions Φ_1 and Φ_2 change by several orders:

1) in the first stage (exponential growth with constant period) Φ_1 and Φ_2 are varied by a factor of approximately 10^{12} compared with the background;

2) in the second stage, connected with the variation of the properties of the system (the coefficients), when the temperature is increased to the critical state the functions Φ_1 and Φ_2 are increased by a factor of 10^2 to 10^3 , and the temperature is increased by $\sim 1000^\circ$;

3) the third stage begins from the critical state, when T continues to increase, and Φ_1 and Φ_2 are decreased by a factor of 10^2 to 10^3 from the value of the function in the critical state;

4) on attaining asymptotically constant subcriticality the functions Φ_1 and Φ_2 continue to decrease with constant period proportional to the value of the subcriticality.

Essentially the second and third stages of the process can be computed; the first and fourth stages are described asymptotically.

2. Computational difficulties

Below it will be convenient for us to write the system of equations (1.1), (1.2), (1.4) in the form

$$Q \partial \Phi / \partial t = L \Phi, \quad (2.1)$$

$$\begin{aligned} \partial T / \partial t &= A_{42} \Phi_2, & \Phi &= (\Phi_1, \Phi_2), \\ L &= \begin{vmatrix} \text{div } D_1 \nabla + A_{11} & \bar{A}_{12} \\ A_{21} & \text{div } D_2 \nabla + A_{22} \end{vmatrix}, & Q &= \begin{vmatrix} 1/v_1 & 0 \\ 0 & 1/v_2 \end{vmatrix}. \end{aligned} \quad (2.2)$$

The correspondence between the "physical" and "mathematical" notations for the coefficients (that is, between the quantities $\Sigma_{a1}, \Sigma_{a2}, \Sigma_{a1}'$ etc and A_{11}, A_{12} etc) is easily established by a simple comparison. We note that the expression

$$A_{12} \Phi_2 + \sum_{i=1}^6 \lambda_i C_i$$

in Eq. (1.1) is written in the form $\bar{A}_{12} \Phi_2$ after using a formula of type (1.10). We do not discuss this in detail since in the calculations considered below the exact calculation of the delayed neutrons is not very important.

A special, fairly complex method was developed and used for calculating the neutron burst; it is described in detail in section 3. Here we will briefly describe formally possible obvious approaches to the problem and estimate the computational difficulties arising in them.

1. The explicit difference scheme:

$$\frac{1}{v_1} \frac{\Phi_1^{n+1} - \Phi_1^n}{\tau} = L_1 \Phi_1^n + A_{11} \Phi_1^n + A_{12} \Phi_2^n, \quad (2.3)$$

$$\frac{1}{v_2} \frac{\Phi_2^{n+1} - \Phi_2^n}{\tau} = L_2 \Phi_2^n + A_{21} \Phi_1^n + A_{22} \Phi_2^n, \quad (2.4)$$

$$\frac{T^{n+1} - T^n}{\tau} = A_{42} \Phi_2^n. \quad (2.5)$$

Here L_1, L_2 are difference operators approximating the corresponding differential operators. Courant's well-known condition imposes a constraint on the step τ . In this case it will be determined by Eq. (2.3); for those values of ν_1 , specified spatial mesh steps and diffusion coefficients used in our calculations, Courant's condition gives $\tau \approx 10^{-7}$ to 10^{-8} . The process must be considered on a time segment $t \approx 1$, that is, the number of steps in the explicit scheme must be $\sim 10^7$, which is completely unrealistic.

2. After replacing (2.3) by a stationary elliptic equation we obtain a computing scheme of the type

$$L_1 \Phi_1^n + A_{11} \Phi_1^n + A_{12} \Phi_2^n = 0, \quad (2.6)$$

the remaining equations are the same as (2.4) and (2.5). In this case Courant's condition is determined by a quantity $\nu_2 \ll \nu_1$, $\tau \approx 10^{-4}$ to 10^{-3} , the number of steps $n \approx 10^4$ to 10^5 . We note that formally each step requires the elliptic equation (2.6) to be solved; in reality, it can obviously be solved not at every time step but only after a large number of explicit steps by the scheme of (2.4); this remark refers to Eq. (2.5) also. Despite this simplification the problem remains too unwieldy.

3. The implicit difference scheme:

$$\begin{aligned} \frac{1}{\nu_1} \frac{\Phi_1^{n+1} - \Phi_1^n}{\tau} &= L_1 \Phi_1^{n+1} + A_{11} \Phi_1^{n+1} + A_{12} \Phi_2^{n+1}, & A_{ij} &= A_{ij}(T^n), \\ \frac{1}{\nu_2} \frac{\Phi_2^{n+1} - \Phi_2^n}{\tau} &= L_2 \Phi_2^{n+1} + A_{21} \Phi_1^{n+1} + A_{22} \Phi_2^{n+1}, \\ \frac{T^{n+1} - T^n}{\tau} &= A_{42} \Phi_2^{n+1}. \end{aligned}$$

In this case there are no constraints on the step τ for stability, but the constraint on τ connected with the approximation accuracy must be taken into account. It can be estimated by considering the solution by the implicit scheme of the simple equation with exponential solution

$$\frac{dx}{dt} = ax, \quad \frac{x^{n+1} - x^n}{\tau} = ax^{n+1}, \quad \text{that is} \quad x^{n+1} = \frac{1}{1 - a\tau} x^n.$$

We compare the numerical solution $x_\tau(t) = x(0) (1 - a\tau)^{-t/\tau}$ with the exact solution $x(t) = x(0) e^{at}$. Making the substitution $(1 - a\tau)^{-1} \simeq e^{a\tau} (1 + a^2 \tau^2 / 2)$ (on the natural assumption that $a\tau \ll 1$), after obvious calculations we obtain

$$x_\tau(t) \simeq x(t) \left(1 + \frac{at a\tau}{2} \right).$$

We have to obtain a solution increasing by a factor of 10^3 (that is, $at = \ln 10^3$) (with the relative accuracy $p\%$, that is, $a\tau/2 = p \cdot 10^{-2}$, from which we obtain for the number of steps

$n=t/\tau$ the estimate $n \sim \ln^2(10^3) \cdot 10^3/2p$. A similar estimate is obtained for calculating the third stage of the process also. Therefore, even satisfying the low relative accuracy of 10%, we obtain for the number of steps the estimate $n \sim 500$; moreover each of them requires a system of elliptic equations to be solved. Using recently developed efficient iterative methods of solving difference elliptic equations and the availability of an excellent initial approximation (Φ^{n+1} differs little from Φ^n), the solution of the problem by the implicit scheme must be regarded as quite feasible on modern computers of the BESM-6 type, but all the same extremely laborious.

4. At first glance the method of alternating directions, not having stability constraints, permits the calculation to be made with the same τ step as in the implicit scheme, but with a considerably simpler algorithm for calculating Φ^{n+1} , without the use of iterative methods. However, a more detailed analysis shows that this is by no means the case.

A numerical experiment was carried out to ascertain the possibilities of this approach. The system of equations

$$\begin{aligned} \frac{1}{10^3} \frac{\partial u}{\partial t} &= \frac{\partial}{\partial x} C \frac{\partial u}{\partial x} + \frac{\partial}{\partial y} C \frac{\partial u}{\partial y} + Av - A_1 u, \\ \frac{\partial v}{\partial t} &= \frac{\partial}{\partial x} D \frac{\partial v}{\partial x} + \frac{\partial}{\partial y} D \frac{\partial v}{\partial y} + Bu - B_1 v, \\ u|_{\Gamma} &= 0, \quad v|_{\Gamma} = 0 \end{aligned}$$

was solved in the square $0 \leq x, y \leq 1$, subdivided into 25 equal parts in each of which the coefficients A, A_1, B, B_1, C, D were constants. The discontinuities in these coefficients (at the boundaries of the parts) were chosen close to the discontinuities in the actual problem. The calculation was performed on a mesh with step $h = 1/60$, and on the whole we can speak of a simplified model of the actual problem.

The standard step of the method, defining the transition from the functions u^n, v^n , relating to the instant t_n , to the functions u^{n+1}, v^{n+1} , consisted of two parts.

The function u^{n+1} is first found from the difference equation approximating the differential equation:

$$(cu_x)_x + (cu_y)_y - A_1 u + Av^n = 0, \quad u|_{\Gamma} = 0.$$

The solution was found by N_u iterations of the method of alternating directions with constant step τ_u , the optimal value of which was calculated by estimating the minimal l and maximal L of the eigenvalues of the elliptic operator: $\tau_u \approx (lL)^{-1/2}$. The function u^n was used as the initial approximation to u^{n+1} .

Then a step of the method of alternating directions is made in the equation

$$v_t = (Dv_x)_x + (Dv_y)_y + Bu^{n+1} - B_1 v, \quad v|_{\Gamma} = 0.$$

In this case τ_v is the time step and the functions u^{n+1}, v^{n+1} relate to the instant $t_{n+1} = t_n + 2\tau_v$. The possibility of solving the problem with a fairly coarse step τ_v was investigated. The calculation was begun with the functions $u^0(x, y) = v^0(x, y) = xy(1-x)(1-y)$, and was continued up to the clear isolation of the exponentially increasing solution $u(t, x, y) = \exp(\lambda t) u^*(x, y)$.

$v(t, x, y) = \exp(\lambda t) v^*(x, y)$. The calculations were protracted (1.5 to 2 hours on the BESM-6 computer programmed in FORTRAN), but the value of λ was determined fairly reliably. The results of the experiment are shown in Table 1. The calculation 1 was performed with a very small value of τ_v , so that the Courant number for the second equation $K_v = 4D\tau_v/h^2 \approx 1$, $N_u = 2$, and this was sufficient, since after one step of the method of alternating directions, that is, after a time $2\tau_v$, the function v was changed by $\sim 0.4\%$. The purpose of this calculation was to obtain the correct value of λ . The results give rise to no doubts; however the solution by a similar method of the actual problem would require $\sim 5 \cdot 10^3$ steps.

TABLE 1

No. of calculation	τ_u	K_u	τ_v	K_v	N_u	λ
1	$0.25 \cdot 10^{-5}$	~ 36	$0.8 \cdot 10^{-4}$	~ 1	2	26.2
2	$0.25 \cdot 10^{-5}$	~ 36	$0.25 \cdot 10^{-2}$	~ 36	10	19.8
3	$0.25 \cdot 10^{-5}$	~ 36	$0.5 \cdot 10^{-2}$	~ 70	10	15.7
4	$0.8 \cdot 10^{-4}$	~ 1000	$0.8 \cdot 10^{-4}$	~ 1	1	26.6

In the following calculations 2 and 3 a considerably greater step in τ_v was used, but this led to considerable errors in the value of λ , although in the functions $u^*(x, y)$, $v^*(x, y)$ and the discrepancies were much smaller. The time step τ_v was not too great, so that $\lambda\tau_v \approx 0.05$ in calculation 2 and ~ 0.1 in calculation 3. Therefore the difference in the values of λ cannot be explained by simple errors of approximation, similar to the error in the integration of the equation $u_t = \lambda u$ by the method of finite differences. Thus, the scheme of first-order accuracy $(u^{n+1} - u^n)/\tau = \lambda u^n$ for $\lambda\tau = 0.1$ gives a solution of the type $\exp(\lambda^* t)$, where $\lambda^* \approx \lambda(1 - \lambda\tau/2)$, that is, an error in λ of about 5% for $\lambda\tau = 0.1$ and 2.5% for $\lambda\tau = 0.05$.

A scheme of second-order accuracy $(u^{n+1} - u^n)/\tau = \lambda(u^{n+1} + u^n)/2$ gives a solution with λ^* differing from λ by $\sim 0.1\%$ for $\lambda\tau = 0.1$. The method of alternating directions, possessing second-order accuracy in τ , is actually close to this scheme. The most probable source of the error is the non-permutability of the operators

$$\frac{\partial}{\partial x} C \frac{\partial}{\partial x}, \quad \frac{\partial}{\partial x} D \frac{\partial}{\partial x}, \quad \frac{\partial}{\partial y} C \frac{\partial}{\partial y}, \quad \frac{\partial}{\partial y} D \frac{\partial}{\partial y}.$$

We illustrate this by the example of the single equation $u_t = (Cu_x)_x + (Cu_y)_y$. Solving this by the method of alternating directions with step τ , we obtain the following connection between $u^{n+1} = u(t_{n+1})$ and $u^n = u(t_n)$:

$$u^{n+1} = \left[E - \tau \frac{\partial}{\partial y} C \frac{\partial}{\partial y} \right]^{-1} \left(E + \tau \frac{\partial}{\partial x} C \frac{\partial}{\partial x} \right) \times \left(E - \tau \frac{\partial}{\partial x} C \frac{\partial}{\partial x} \right)^{-1} \left(E + \tau \frac{\partial}{\partial y} C \frac{\partial}{\partial y} \right) u^n = B_\tau u^n.$$

It is known that a lengthy calculation by this formula leads to the isolation of the principal eigenfunction of the operator B_τ . Because of the non-permutability of the operators, the eigenfunctions of B_τ are not identical with the eigenfunctions of the operator

$$\frac{\partial}{\partial x} C \frac{\partial}{\partial x} + \frac{\partial}{\partial y} C \frac{\partial}{\partial y};$$

this difference is the more important the greater the value of τ .

3. The method of solution

The first stage of the physical process proceeds at the fixed temperature $T(r, z, t)$, that is, with values of the coefficients of the system constant in time. Then the solution can be represented by a Fourier series:

$$\Phi(r, z, t) = \sum_{k=1}^{\infty} c_k \exp(\lambda_k t) \psi_k(r, z); \quad (3.1)$$

here $\psi_k(r, z)$ are the eigenfunctions, and λ_k are the eigenvalues of the operator $L\psi_k = \lambda_k Q\psi_k$. The operator L is elliptic, its spectrum is real and overbounded: $-\infty < \dots < \lambda_k < \dots < \lambda_1 = \Lambda$. If

$\Lambda > 0$ the system is supercritical, if $\Lambda < 0$ the system is subcritical. At the initial instant ($t = 0$) a supercritical is created in the system.

We write (3.1) in the form

$$\Phi(r, z, t) = \exp(\lambda_1 t) \sum_{k=1}^{\infty} c_k \exp[-(\lambda_1 - \lambda_k)t] \psi_k(r, z), \quad (3.2)$$

$$(\lambda_1 - \lambda_k)t \gg 1, \quad \text{if } k \neq 1.$$

It is obvious from (3.2) that at the first stage (since its duration is very great) a dominant part in the right side of (3.1) will be played by only the term corresponding to the extreme right point of the spectrum: $\Phi(r, z, t) \approx c_1 \exp(\lambda_1 t) \psi_1(r, z)$.

Therefore, instead of calculating the first stage of the process we must determine the first eigenfunction $\psi_1(r, z)$ of the operator $L\psi = \lambda Q\psi$, corresponding to the extreme right point of the spectrum $\Lambda = \lambda_1$, and proceed to the calculation of the second stage of the process, when the temperature begins to change. As initial data for calculating the second stage we have to take the function $\Phi(r, z, t_0) = N(t_0) \psi_1(r, z)$, where $N(t_0)$ is a value sufficiently great for the process of temperature variation to have begun. The quantity $N(t_0)$ can be estimated: the characteristic time of the process Δt (0.1–1 sec) is known, the characteristic variation of temperature is known ΔT ($\sim 1000^\circ$). We can use the tentative relation $\Delta T / \Delta t \approx A_{42} N(t_0) \max \varphi_2(r, z)$, then $N(t_0) = \Delta T (A_{42} \max \varphi_2 \Delta t)^{-1}$, where $\psi_1 = (\varphi_1, \varphi_2)$.

Usually $N(t_0)$ is taken less than the required value by a factor of 10 to 100, this leads to a situation where the calculation of the first few steps in time is performed with an actually unchanged $T(r, z, t)$, that is, the "tail" of the first stage of the process is computed (see Fig. 1).

As for the initial "background" $\Phi(r, z, 0)$, the following natural assumptions about it are made:

1) in the expansion of $\Phi(r, z, 0)$ in a Fourier series the coefficient of the first eigenfunction is not too small in comparison with the others;

2) the absolute value of $\Phi(r, z, 0)$ is sufficiently small and the time necessary for $\|\Phi(r, z, t)\|$ to become a quantity of order $\sim N(t_0)$ at which the change of temperature has begun, is sufficiently great for all the terms, except the first, in the sum of (3.1) to be neglected.

For the calculation of the second and third stages of the process, when the temperature effects lead to a spatially-inhomogeneous variation of the sections, a method was used which is a generalization of the Fourier method, perhaps rather less accurate, but more economical from the point of view of computing time.

An approximate solution $\Phi(r, z, t) = (\Phi_1(r, z, t), \Phi_2(r, z, t))$ of the system (2.1), (2.2) is sought in the form

$$\Phi(r, z, t) = N(t_0) \exp \left[\int_{t_0}^t \Lambda(\tau) d\tau \right] \varphi(r, z, t), \quad (3.3)$$

where $\varphi(r, z, t) = (\varphi_1(r, z, t), \varphi_2(r, z, t))$ is the normed first eigenfunction, and $\Lambda(t)$ is the first eigenvalue of the operator

$$L(t)\varphi(r, z, t) = \Lambda(t)Q(t)\varphi(r, z, t); \quad (3.4)$$

L, Λ, Q depend on the time implicitly via the temperature $T(r, z, t)$ occurring in the coefficients.

The equations for $C_i(r, z, t)$ and $T(r, z, t)$ assume the form

$$\begin{aligned} \frac{\partial C_i}{\partial t} &= -\lambda_i C_i + \beta_i A_{32} N(t_0) \exp \left[\int_{t_0}^t \Lambda(\tau) d\tau \right] \varphi_2(r, z, t), \\ \frac{\partial T}{\partial t} &= A_{42} N(t_0) \exp \left[\int_{t_0}^t \Lambda(\tau) d\tau \right] \varphi_2(r, z, t). \end{aligned} \quad (3.5)$$

To estimate the error of this method we substitute the approximate solution (3.3) into the left side of (2.1); we obtain

$$\begin{aligned} Q \frac{\partial \Phi}{\partial t} &= Q \frac{\partial}{\partial t} \left\{ N(t_0) \exp \left[\int_{t_0}^t \Lambda(\tau) d\tau \right] \varphi(r, z, t) \right\} \\ &= Q \left\{ \Lambda(t) N(t_0) \exp \left[\int_{t_0}^t \Lambda(\tau) d\tau \right] \varphi + N(t_0) \exp \left[\int_{t_0}^t \Lambda(\tau) d\tau \right] \frac{\partial \varphi}{\partial t} \right\} \\ &= Q \left\{ \Lambda(t) \Phi + N(t_0) \exp \left[\int_{t_0}^t \Lambda(\tau) d\tau \right] \frac{\partial \varphi}{\partial t} \right\} = L\Phi + Q(\Delta\Phi), \\ \Delta\Phi &= \left(\frac{\partial \ln \varphi_1}{\partial t} \Phi_1, \frac{\partial \ln \varphi_2}{\partial t} \Phi_2 \right), \\ \frac{\partial \Phi_i}{\partial t} &= \left(\Lambda(t) + \frac{\partial \ln \varphi_i}{\partial t} \right) \Phi_i, \quad i=1, 2, \end{aligned} \quad (3.6)$$

that is, (3.3) satisfies not (2.1), but the equation

$$Q \left(\frac{\partial \Phi}{\partial t} - \Delta \Phi \right) = L \Phi. \quad (3.7)$$

It is obvious from (3.6) that the effect of the error depends on the quantities

$$\frac{1}{v_i} \frac{\partial \ln \varphi_i}{\partial t} \Phi_i, \quad i=1, 2. \quad (3.8)$$

Calculations have shown that the components of (3.8) are less by a factor of more than 100 than the individual terms occurring in the expression $Q \partial \Phi / \partial t - L \Phi$.

In connection with Eq. (3.7) it is also understood that the method of norming the eigenfunction still remains arbitrary. The normalization is defined by the relation

$$\int_0^H \int_0^R [\varphi_2(r, z, t)]^2 r dr dz = 1. \quad (3.9)$$

With this normalization the value of $\partial \varphi_2 / \partial t$ is a minimum; the value of $\partial \varphi_1 / \partial t$ is greater than, for example, in a normalization of the type

$$\|\varphi\|^2 = \int_0^H \int_0^R (\varphi_1^2 + \varphi_2^2) r dr dz = 1,$$

but it is the smallness of $\partial \varphi_2 / \partial t$, which is important to us, since the equation contains the expressions

$$\frac{1}{v_1} \frac{\partial \varphi_1}{\partial t}, \quad \frac{1}{v_2} \frac{\partial \varphi_2}{\partial t}, \quad \text{and} \quad v_1 \approx 10^3 v_2.$$

Therefore, the solution of the system (3.3)–(3.5) satisfies Eqs. (1.3), it satisfies with high accuracy Eqs. (1.1) and (1.2), it satisfies the boundary conditions (1.8), (1.9), and also the initial conditions, since as a result of the first stage of the process up to the beginning of the second stage in the solution of (3.1) an overwhelming part is played by the first term, which may be written in the form $N(t_0) \varphi(r, z, t_0)$.

The basis of the numerical method is to find the first eigenfunction (φ_1, φ_2) and the corresponding eigenvalue $\Lambda(t)$ of the elliptic operator $Q^{-1}L(t)$ at each time step. The calculation of the process usually consists of 20 to 30 steps.

We describe the structure of one time step, ignoring for brevity the effect of the delayed neutrons.

The time t is subdivided into intervals small relative to the rate of change of temperature, called "steps" in t : $t_0 < t_1 < \dots$, however, relative to the speed of the neutron processes the step $\Delta t_n = t_{n+1} - t_n$ is "large".

At the instant t_n let us have $T(r, z, t_n)$, $N(t_n)$, $\Lambda(t_n)$, $\varphi_1(r, z, t_n)$, $\varphi_2(r, z, t_n)$. One step, the transition from t_n to t_{n+1} consists of the following stages:

1) the choice of the step Δt_n from the calculation of the given (~ 10 to 15%) increment of temperature;

2) the calculation of the temperature $T(r, z, t_{n+1})$ at the instant $t_{n+1} = t_n + \Delta t_n$:

$$T(r, z, t_{n+1}) = T(r, z, t_n) + \int_{t_n}^{t_{n+1}} A_{42}(r, z, t_n) \varphi_2(r, z, t_n) N(\tau) d\tau;$$

at the same time we compute

$$N(t_{n+1}) = N(t_n) \exp \left[\int_{t_n}^{t_{n+1}} \Lambda(\tau) d\tau \right];$$

3) finding the first eigenfunction $(\varphi_1(r, z, t_{n+1}), \varphi_2(r, z, t_{n+1}))$ and eigenvalue $\Lambda(t_{n+1})$ of the operator

$$L(t_{n+1})\varphi = \Lambda(t_{n+1})Q(t_{n+1})\varphi.$$

The description of the method of finding the first eigenfunction forms the subject of a separate paper (see [4, 5]).

This completes the calculation of the fundamental values at one time step; for the next time step we again have $T(r, z, t_{n+1})$, $N(t_{n+1})$, $\Lambda(t_{n+1})$, $\varphi_1(r, z, t_{n+1})$, $\varphi_2(r, z, t_{n+1})$.

Remark. The reasons for the leading role of the first eigenfunction in such phenomena is well known to physicists, and we are informed by Ya. B. Zel'dovich and V. Ya. Gol'din, that in their time they were used in some calculations.

4. Examples of calculations

The method explained in section 3 was used to simulate on the BESM-6 computer the kinetics of a neutron burst in a pulsed reactor of heat-capacity type IGR [3].

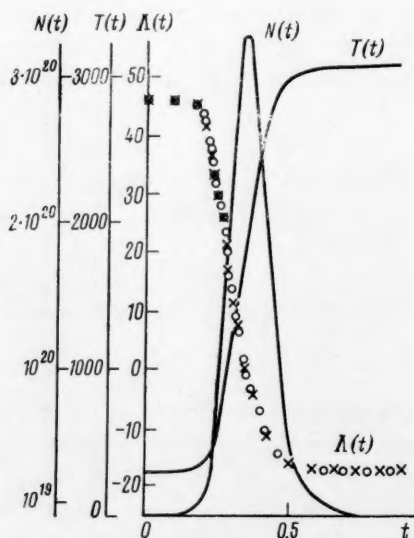


FIG. 1.

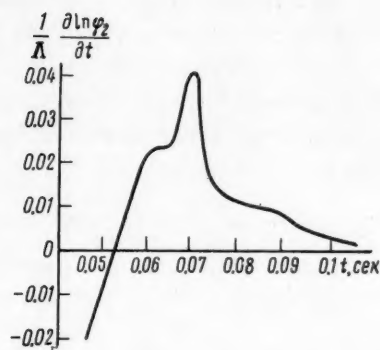


FIG. 2.

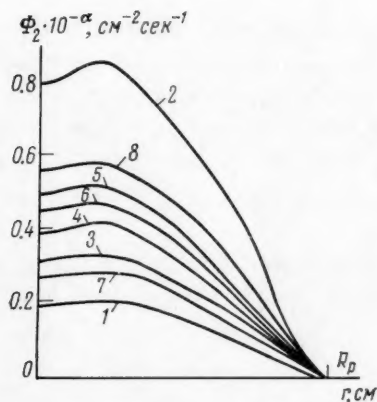


FIG. 3.

The function $\Phi_2 \cdot 10^{-\alpha}$: 1 is for $t = 0.0407$ sec, 2 is for $t = 0.0477$ sec, 3 is for $t = 0.059$ sec, 4 is for $t = 0.0619$ sec, 5 is for $t = 0.0705$ sec, 6 is for $t = 0.0769$ sec, 7 is for $t = 0.088$ sec, 8 is for $t = 0.112$ sec; $\alpha = 18$ for 1, 2, 8; $\alpha = 19$ for 3 - 7.

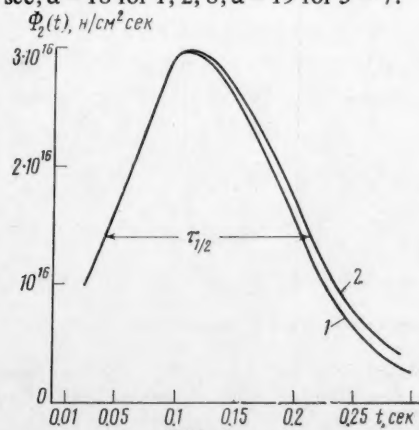


FIG. 4.

Figure 1 shows the functions $\Lambda(t)$, $N(t)$ and $T(t) = \max_{(r,z)} T(r, z, t)$, characteristic for the pulsed mode. All four stages of the phenomenon considered are easily seen. As a control this pulse was calculated twice; in the first case the step Δt_n was chosen so as to ensure an increment of temperature $T(r, z, t_n)$ of $\sim 10\%$, the corresponding values of $\Lambda(t)$ are shown by dots on the graph. Then the calculation was repeated with a temperature increment of $\sim 20\%$ per step, the corresponding values of $\Lambda(t)$ are shown by circles. In both cases all the values of $\Lambda(t)$ are shown on the graph, and this permits us to judge the number of time steps. We must bear in mind that the function $\Phi(t) = N(t) \varphi$ has physical significance and for $N(t) \approx 3 \cdot 10^{20}$ the quantity $\Phi \approx 10^{18} \text{ cm}^{-2} \text{ sec}^{-1}$.

Figure 2 gives an idea of the accuracy of the approximation (13), that is, of the possibility of neglecting the quantity $\partial \ln \varphi / \partial t$. The graph shows the function

$$\frac{\partial \ln \varphi_2(t)}{\partial t} \Lambda^{-1}(t), \quad \text{where} \quad \varphi_2(t) = \max_{(r,z)} \varphi_2(r, z, t).$$

This relation is taken from the calculation of a very fast pulse with great initial supercriticality $\Lambda(0) = 200 \text{ sec}^{-1}$.

In the simulation of a neutron burst into instantaneous neutrons the calculation of the second stage of the process began with the level $\Phi_0 = 10^{15} \text{ n/cm}^2 \text{ sec}$, the initial supercriticality $\Lambda(0) = 201 \text{ sec}^{-1}$, the time for reaching criticality of the reactor $t \approx 0.07 \text{ sec}$ ($\Lambda = 0$), the total duration of the pulse $\sim 0.1 \text{ sec}$, and halfwidth of the pulse $\tau_{1/2} \approx 0.045 \text{ sec}$. At the end of the burst, when the mean temperature of the stack of the active zone $T = 2430^\circ \text{K}$, the value of the subcriticality $\Lambda \approx -60 \text{ sec}^{-1}$. This pulse is shown in Fig. 3.

Figure 4 illustrates the effect of the delayed neutrons, which is shown by calculations to be appreciable only in slow bursts, described by small values of the initial supercriticality, $\Lambda(0) \approx 20 \text{ sec}^{-1}$. The function shown

$$\Phi_2(t) = \max_{(r,z)} \Phi_2(r, z, t),$$

obtained without allowing for the delayed neutrons (curve 1, $\tau_{1/2} = 0.2 \text{ sec}$) and allowing for the delayed neutrons (curve 2, $\tau_{1/2} = 0.205 \text{ sec}$). It is seen that allowing for the integral term (1.10) in Eq. (1.1) gives a maximum difference in the amplitude of the neutron flux of the order of several percent, the pulse halfwidth $\tau_{1/2}$ is then increased by approximately 2–3%. For these calculations the function $\Lambda^{-1} \partial \ln \varphi_2 / \partial t$ is of approximately the same nature as the function of Fig. 2.

In conclusion we mention the paper [6] in which a method of numerical integration of the one-dimensional kinetic equation with a sharply increasing solution is proposed. The principal content of this paper is the surmounting of the constraints on the time step.

Translated by J. Berry

REFERENCES

1. AKCASY, A. Z. *Mathematical models in nuclear reactor dynamics*. Acad. Press, New York, 1972.
2. STACEY, W. M., Jr. *Space-time nuclear reactor kinetics*, Acad. Press, New York, 1969.
3. KURCHATOV, I. V. et al. The pulsed graphite reactor IGR. The Third Geneva Conference on the Peaceful Uses of Atomic Energy. Report No. 322 (III Zhenevskaya konf. mirnomu ispol'zovaniya at. energii. Dokl. No. 322), MAGATE, Wien, 1964.

4. KLIMOV, A. D., STRAKHOVSKAYA, L. G., FEDORENKO, R. P. and CHIKHLADZE, I. L. A method of calculating the space-time kinetics of a pulsed reactor in $(r - z)$ -geometry. Preprint IPM, No. 42, Akad. Nauk SSSR, 1971.
5. STRAKHOVSKAYA, L. G. An iterative method of finding the principal eigenfunction of a difference elliptic operator. Preprint No. 77, IPM Akad. Nauk SSSR, 1972.
6. GOL'DIN, V. Ya., DANILOVA, G. V. and CHETVERUSHKIN, B. N. An approximate method of calculating the non-stationary kinetic equation. In: *Computational methods in transfer theory* (Vychisl. metody v teorii perenosa), 50-58, Atomizdat, Moscow, 1969.

ANALYSE ASYMPTOTIQUE DES ECOULEMENTS DE FLUIDES VISQUEUX COMPRESSIBLES A FAIBLE NOMBRE DE MACH*

I. Cas des fluides non pesants

R. Kh. ZEYTOUNIAN

Lille, France

(Received 10 November 1975)

THE GENERAL problem of the analysis of a stationary sink of some viscous compressible medium for low Mach numbers is studied by means of the asymptotic expansion of the solution in powers of the characteristic Mach number M_∞ . It is shown that there may be various forms of these expansions depending on the temperature conditions on a closed surface bounding the given sink.

On considère le problème général de l'analyse d'un écoulement stationnaire d'un fluide visqueux compressible à faible nombre de Mach par une méthode de perturbation, dans laquelle la solution est représentée par des développements asymptotiques par rapport à un nombre de Mach caractéristique M_∞ . On montre que ces développements peuvent prendre différentes formes en fonction de la condition pour la température sur la surface fermée Σ^* qui délimite intérieurement l'écoulement. De manière précise, si cette dernière condition est écrite sous la forme:

$$T = T_\infty^* + \Delta T_0^* \Xi, \text{ sur } \Sigma^*, \text{ avec } T_\infty^* \text{ et } \Delta T_0^*$$

des températures caractéristiques constantes liées respectivement à l'écoulement uniforme loin de Σ^* et à la variation de température Ξ sur Σ^* , et que l'on suppose que $\tau_0 = \Delta T_0^* / T_\infty^* \rightarrow 0$, avec $M_\infty \rightarrow 0$, de telle façon que: $\tau_0 = \Lambda_0 M_\infty^\omega$, où Λ_0 est un paramètre de similitude constant et $\omega > 0$ un nombre réel supposé donné, alors il se présente trois cas: $\omega < 2$, $\omega = 2$ et $\omega > 2$.

Les cas $\omega < 2$ et $\omega > 2$ conduisent à des développements asymptotiques qui sont définis à partir de la séquence asymptotique $M_\infty^{2p + \omega q}$ avec $p, q = 0, 1, \dots$; lorsque $\omega = 2$ la séquence asymptotique s'identifie avec celle qui est classiquement connue M_∞^{2n} , avec $n \equiv p + q = 0, 1, \dots$ (celle de Janzen-Rayleigh [1, 2]).

On précise ainsi, en particulier, l'évolution de la température et de la masse volumique dans un écoulement à faible nombre de Mach (écoulement quasi-incompressible); de ce fait, chaque fois que l'on peut calculer un écoulement incompressible de fluide visqueux on peut aussi lui associer un calcul des champs de la température et de la masse volumique et obtenir ainsi une représentation correcte (au sens des développements asymptotiques) de la solution des équations de Navier-Stokes.

*Zh. vychisl. Mat. mat. Fiz., 17, 175-182, 1977.

1. Formulation de problème

Le fluide visqueux compressible est supposé être un gaz parfait à chaleurs spécifiques c_p et c_v constantes; il s'étend à l'infini dans toutes les directions et il est limité intérieurement par la surface fermée Σ^* . A de grandes distances de Σ^* on suppose qu'il existe un écoulement uniforme de vitesse v_∞^* constante dans lequel la pression p_∞^* , la masse volumique ρ_∞^* et la température T_∞^* sont constantes. On utilise un repère et un système de coordonnées cartésiennes orthogonales $\{0, x_k^*\}$.

Enfin, on désigne par v_k^* , p^* , ρ^* et T^* les composantes de la vitesse, la pression, la masse volumique et la température dans l'écoulement induit par la présence de Σ^* au sein de l'écoulement uniforme.

Dans tout ce qui suit la vitesse de l'écoulement en tout point est supposé très petite devant la célérité locale du son; en d'autres termes, le nombre de Mach caractéristique de l'écoulement:

$$M_\infty = \frac{U_\infty^*}{(\gamma R T_\infty^*)^{1/2}} \ll 1, \quad (1.1)$$

où U_∞^* est une vitesse caractéristique liée à v_∞^* et $\gamma = c_p/c_v$, $R = c_p(\gamma - 1)/\gamma$. Par la suite la solution des équations de Navier - Stokes est donc représentée par des développements asymptotiques par rapport au nombre de Mach caractéristique M_∞ . Notons que pour les écoulements de fluides parfaits compressibles c'est une méthode classique dans le cas d'écoulements stationnaires connue sous le nom de méthode de Janzen-Rayleigh [1, 2].

Par contre le cas instationnaire présente une différence fondamentale par rapport au cas stationnaire, car on montre [3, 4] que ces développements, qui sont en fait des développements intérieurs [5], ne sont pas alors uniformément valables à grande distance de Σ^* . De ce fait la solution à grande distance de Σ^* est représentée par des développements extérieurs. Les conditions de raccord entre ces deux développements permettant, ensuite, de déterminer complètement la solution aussi bien dans le domaine distal que dans le domaine proximal et ce en écrivant le développement composite à l'ordre d'approximation correspondant. En particulier, il s'avère qu'il est nécessaire d'introduire dans les développements inférieurs des termes en puissance impaires de M_∞ (termes qui sont absents dans le développement classique de Janzen-Rayleigh) pour que ce raccordement soit possible.

En ce qui nous concerne ici nous nous intéressons plus particulièrement au fluide visqueux compressible et au domaine proximal proche de la surface Σ^* sur laquelle il faut écrire les conditions d'adhérence du fluide visqueux et une condition pour la température. A cet effet, et pour éviter toute ambiguïté, nous supposons que l'écoulement est stationnaire; les équations de Navier-Stokes qui régissent cet écoulement stationnaire de gaz parfait s'écriront, sous forme adimensionnelle,

$$v_k \frac{\partial \rho}{\partial x_k} + \rho \frac{\partial v_k}{\partial x_k} = 0, \quad (1.2a)$$

$$\begin{aligned}
 \rho v_k \frac{\partial v_i}{\partial x_k} + \frac{1}{\gamma M_\infty^2} \frac{\partial p}{\partial x_i} &= \frac{1}{\text{Re}} \left[\frac{\partial^2 v_i}{\partial x_k^2} + \frac{1}{3} \frac{\partial}{\partial x_i} \left(\frac{\partial v_k}{\partial x_k} \right) \right], \\
 v_k \frac{\partial}{\partial x_k} \left(T - \frac{\gamma-1}{\gamma} p \right) &= \frac{1}{\text{Pr}} \frac{1}{\text{Re}} \frac{\partial^2 T}{\partial x_k^2}, \\
 -(\gamma-1) \frac{M_\infty^2}{\text{Re}} \left[\frac{2}{3} \left(\frac{\partial v_k}{\partial x_k} \right)^2 - \frac{1}{2} \left(\frac{\partial v_i}{\partial x_k} + \frac{\partial v_k}{\partial x_i} \right)^2 \right], \\
 p &= \rho T.
 \end{aligned} \tag{1.2}$$

Les quantités dimensionnelles étant caractérisées par un astérisque, les variables sans dimensions qui apparaissent dans ces équations (1.2) sont définies de la façon suivante: $\rho = \rho^*/\rho_\infty$, $p = p^*/p_\infty$, $T = T^*/T_\infty$, $v_k = v_k^*/U_\infty$, $x_k = x_k^*/L_0$ où L_0 est une longueur caractéristique (liée à Σ^* , en particulier). En plus du nombre de Mach caractéristique (1.1) il s'introduit dans les équations adimensionnelles (1.2) le nombre de Reynolds $\text{Re} = U_\infty^* L_0^*/\nu_0^*$, où ν_0^* est le coefficient (constant) cinématique de viscosité et le nombre de Prandtl $\text{Pr} = c_p \nu_0^* \rho_\infty/k_0^*$ où k_0^* est le coefficient (constant) de conduction. Notons enfin que les équations (1.2) sont écrites sous l'hypothèse de Stokes (coefficient de viscosité volumique nulle).

La solution des équations (1.2) satisfait aux conditions aux limites suivantes (l'écoulement étant supposé continu partout):

- 1) conditions à l'infini où il existe un écoulement uniforme,
- 2) conditions d'adhérence sur Σ : $v_k = 0$;
- 3) condition pour la température sur Σ , que l'on écrira sous la forme adimensionnelle suivante:

$$T = 1 + \tau_0 \Xi, \tag{1.3}$$

où $\tau_0 = \Delta T_0^*/T_\infty^*$, avec ΔT_0^* une variation caractéristique de température liée à la fonction Ξ qui est supposée connue.

Les conditions ci-dessus sont supposées suffisantes pour rendre la solution des équations (1.2) unique; de plus, on se donne la position et la forme de Σ^* .

2. Ecoulement incompressible

Représentons la solution des équations (1.2) satisfaisant aux conditions aux limites, lorsque $M_\infty \rightarrow 0$ à x_k Pr et Re fixés, par des développements asymptotiques de la forme:

$$\begin{aligned}
 v_k &= v_k^{(0)} + M_\infty^\alpha v_k^{(\alpha)} + \dots, & p &= p^{(0)} + M_\infty^\beta p^{(\beta)} + \dots, \\
 \rho &= \rho^{(0)} + M_\infty^\gamma \rho^{(\gamma)} + \dots, & T &= T^{(0)} + M_\infty^\delta T^{(\delta)} + \dots,
 \end{aligned}$$

où α, β, γ et δ sont des nombres réels positifs pour l'instant arbitraires.

La seconde équation du système (1.2) implique nécessairement que

$$\partial p^{(0)} / \partial x_i = 0 \Rightarrow \beta = 2$$

soit, en tenant compte, pour $p^{(0)}$, des conditions à l'infini

$$p^{(0)} \equiv 1.$$

Les équations (1.2) peuvent alors s'écrire, en première approximation, sous la forme:

$$\begin{aligned} v_k^{(0)} \frac{\partial \rho^{(0)}}{\partial x_k} + \rho^{(0)} \frac{\partial v_k^{(0)}}{\partial x_k} &= 0, \\ \rho^{(0)} v_k^{(0)} \frac{\partial v_i^{(0)}}{\partial x_k} + \frac{1}{\gamma} \frac{\partial p^{(2)}}{\partial x_i} &= \frac{1}{\text{Re}} \left[\frac{\partial^2 v_i^{(0)}}{\partial x_k^2} + \frac{1}{3} \frac{\partial}{\partial x_i} \left(\frac{\partial v_k^{(0)}}{\partial x_k} \right) \right], \\ \rho^{(0)} v_k^{(0)} \frac{\partial T^{(0)}}{\partial x_k} &= \frac{1}{\text{Pr}} \frac{1}{\text{Re}} \frac{\partial^2 T^{(0)}}{\partial x_k^2}, \\ \rho^{(0)} T^{(0)} &= 1. \end{aligned} \quad (2.1)$$

Au niveau du système limite (2.1) les conditions d'adhérence et celles à l'infini restent inchangées (puisque nous sommes en écoulement stationnaire, régulier à l'infini). Si, d'autre part, l'on ne fait aucune hypothèse sur le paramètre τ_0 , au niveau de la condition (1.3), alors on ne pourra effectuer aucune simplification supplémentaire au niveau du système limite (2.1).

Pour obtenir les équations régissant l'écoulement incompressible il faut admettre que: dans la condition à la limite (1.3) le paramètre τ_0 est beaucoup plus petit que l'unité

$$\tau_0 \ll 1 \Rightarrow \Delta T_0^* \ll T_\infty^*;$$

de manière précise on admettra que: $\tau_0 \rightarrow 0$ avec $M_\infty \rightarrow 0$ de telle façon que la relation de similitude

$$\tau_0 = \Lambda_0 M_\infty^\omega \quad (2.2)$$

soit satisfaite, avec Λ_0 un paramètre de similitude constant et $\omega > 0$ un nombre réel supposé donné.

Dans ce cas et grâce à l'hypothèse (2.2) on peut associer aux équations (2.1) la condition:

$$T^{(0)} = 1, \text{ sur } \Sigma.$$

Puisque $T^{(0)} = 1$ et $\rho^{(0)} = 1$ à l'infini il est clair que les équations (2.1) admettent la solution triviale:

$$T^{(0)} \equiv 1, \quad \rho^{(0)} \equiv 1$$

ce qui nous conduit aux équations de Navier pour un écoulement incompressible [6]

$$\frac{\partial v_k^{(0)}}{\partial x_k} = 0, \quad v_k^{(0)} \frac{\partial v_i^{(0)}}{\partial x_k} = -\frac{1}{\gamma} \frac{\partial p^{(2)}}{\partial x_i} + \frac{1}{\text{Re}} \frac{\partial^2 v_i^{(0)}}{\partial x_k^2}, \quad (2.3)$$

auxquelles il faut associer les conditions

$$\begin{aligned} v_k^{(0)} &= 0, \quad \text{sur } \Sigma, \\ v_k^{(0)} &\rightarrow v_{\infty, k} \text{ et } p^{(2)} \rightarrow 0, \text{ à l'infini.} \end{aligned} \quad (2.4)$$

3. Ecoulement quasi-incompressible

Au niveau de la première approximation, qui conduit à l'écoulement incompressible, on perd toute information sur l'évolution de la température et de la masse volumique dans l'écoulement à faible nombre de Mach.

Considérons, tout d'abord l'équation de continuité; en tenant compte de ce que $\rho^{(0)} \equiv 1$ et $\partial v_k^{(0)} / \partial x_k = 0$ on obtient pour $\rho^{(\gamma)}$ l'équation

$$v_k^{(0)} \frac{\partial \rho^{(\gamma)}}{\partial x_k} + M_\infty^{\alpha-\gamma} \frac{\partial v_k^{(\alpha)}}{\partial x_k} = 0.$$

seul le cas $\alpha = \gamma$ conduit à une dégénérescence significative [7], ce qui donne:

$$v_k^{(0)} \frac{\partial \rho^{(\gamma)}}{\partial x_k} + \frac{\partial v_k^{(\gamma)}}{\partial x_k} = 0,$$

la valeur $\gamma > 0$, restant à ce stade indéterminée.

Afin de préciser la valeur de γ considérons la loi d'état; il peut alors se présenter trois cas:

$$2 = \delta = \gamma, \delta = \gamma \text{ et } \gamma = 2.$$

Enfin, l'équation de l'énergie peut s'écrire sous la forme suivante:

$$\begin{aligned} v_k^{(0)} \frac{\partial T^{(\delta)}}{\partial x_k} - \frac{\gamma-1}{\gamma} M_\infty^{2-\delta} v_k^{(0)} \frac{\partial p^{(2)}}{\partial x_k} &= \frac{1}{Pr} \frac{1}{Re} \frac{\partial^2 T^{(\delta)}}{\partial x_k^2} \\ &+ \frac{\gamma-1}{\gamma} M_\infty^{2-\delta} \left(\frac{\partial v_i^{(0)}}{\partial x_k} + \frac{\partial v_k^{(0)}}{\partial x_i} \right)^2 \end{aligned}$$

à laquelle, sous l'hypothèse (2.2), on doit associer la condition

$$T^{(\delta)} = M_0^{\omega-\delta} \Lambda_0 \Xi, \text{ sur } \Sigma. \quad (3.1)$$

Afin de ne pas perdre la condition (3.1) «il semble», à première vue, que l'on doit admettre que

$$\delta = \omega, \text{ avec } \omega > 0 \text{ un réel donné.} \quad (3.2)$$

Par la suite il faut considérer trois cas

$$1. 0 < \omega < 2$$

Dans ce cas, si l'on admet que (3.2) a effectivement lieu, on obtient que:

$$0 < \alpha = \gamma = \delta = \omega < 2$$

et les fonctions $\rho^{(\omega)}$, $v_k^{(\omega)}$ et $T^{(\omega)}$ satisfont aux équations:

$$\frac{\partial v_k^{(\omega)}}{\partial x_k} = -v_k^{(0)} \frac{\partial \rho^{(\omega)}}{\partial x_k}, \quad (3.3a)$$

$$\rho^{(\omega)} = -T^{(\omega)}, \quad (3.3b)$$

$$v_k^{(0)} \frac{\partial T^{(\omega)}}{\partial x_k} = \frac{1}{\text{Pr}} \frac{1}{\text{Re}} \frac{\partial^2 T^{(\omega)}}{\partial x_k^2}. \quad (3.3c)$$

Pour obtenir l'équation du mouvement associé au système (3.3), il faut que le développement asymptotique de la pression soit de la forme

$$p = 1 + M_\infty^2 p^{(2)} + M_\infty^{2+\omega} p^{(2+\omega)} + \dots$$

ce qui donne l'équation de mouvement suivante:

$$\begin{aligned} & \frac{\partial v_i^{(0)}}{\partial x_k} v_k^{(\omega)} + v_k^{(0)} \frac{\partial v_i^{(\omega)}}{\partial x_k} + \frac{1}{\gamma} \frac{\partial p^{(2+\omega)}}{\partial x_i} \\ &= \frac{1}{\text{Re}} \left\{ \frac{\partial^2 v_i^{(\omega)}}{\partial x_k^2} + \frac{1}{3} \frac{\partial}{\partial x_i} \left(\frac{\partial v_k^{(\omega)}}{\partial x_k} \right) \right\} + T^{(\omega)} v_k^{(0)} \frac{\partial v_i^{(0)}}{\partial x_k}. \end{aligned} \quad (3.4)$$

On détermine, tout d'abord, $T^{(\omega)}$ à partir de l'équation linéaire homogène (3.3c) à laquelle il faut associer les conditions:

$$T^{(\omega)} = \Lambda_0 \Xi, \text{ sur } \Sigma \text{ et } T^{(\omega)} \rightarrow 0, \text{ à l'infini.} \quad (3.5)$$

Puis $\rho^{(\omega)} = -T^{(\omega)}$ et ensuite $v_k^{(\omega)}$ et $p^{(2+\omega)}$ à partir du système des deux équations linéaires non homogène (3.3a) et (3.4) auquel il faut associer les conditions:

$$v_k^{(\omega)} = 0, \text{ sur } \Sigma; \quad v_k^{(\omega)} \text{ et } p^{(2+\omega)} \rightarrow 0, \text{ à l'infini.} \quad (3.6)$$

$$2. \quad \omega = 2$$

Dans ce cas

$$T = 1 + M_\infty^2 \Lambda_0 \Xi, \text{ sur } \Sigma$$

et on retombe sur le cas classique de Janzen et Rayleigh [1, 2]:

$$0 < \alpha = \gamma = \delta = \omega = 2.$$

Une fois de plus $v_k^{(0)}$ et $p^{(2)}$ satisfont au problème de l'hydrodynamique classique (2.3), (2.4). Quant aux fonctions $v_k^{(2)}$, $p^{(4)}$, $\rho^{(2)}$ et $T^{(2)}$ elles doivent être déterminées à partir du système d'équations:

$$T^{(2)} + \rho^{(2)} = p^{(2)}; \quad (3.7a)$$

$$\begin{aligned} v_k^{(0)} \frac{\partial T^{(2)}}{\partial x_k} &= \frac{1}{\text{Pr}} \frac{1}{\text{Re}} \frac{\partial^2 T^{(2)}}{\partial x_k^2} + \frac{\gamma-1}{2 \text{Re}} \left(\frac{\partial v_i^{(0)}}{\partial x_k} + \frac{\partial v_k^{(0)}}{\partial x_i} \right)^2 \\ &+ \frac{\gamma-1}{\gamma} v_k^{(0)} \frac{\partial p^{(2)}}{\partial x_k}; \end{aligned} \quad (3.7b)$$

$$\frac{\partial v_k^{(2)}}{\partial x_k} = -v_k^{(0)} \frac{\partial \rho^{(2)}}{\partial x_k}; \quad (3.7c)$$

$$\begin{aligned} & \frac{\partial v_i^{(0)}}{\partial x_k} v_k^{(2)} + v_k^{(0)} \frac{\partial v_k^{(2)}}{\partial x_k} + \frac{1}{\gamma} \frac{\partial p^{(4)}}{\partial x_i} \\ &= \frac{1}{\text{Re}} \left\{ \frac{\partial^2 v_i^{(2)}}{\partial x_k^2} + \frac{1}{3} \frac{\partial}{\partial x_i} \left(\frac{\partial v_k^{(2)}}{\partial x_k} \right) \right\} - \rho^{(2)} v_k^{(0)} \frac{\partial v_i^{(0)}}{\partial x_k}. \end{aligned} \quad (3.7d)$$

On détermine $T^{(2)}$ à partir de l'équation linéaire non homogène (3.7b) avec les conditions:

$$T^{(2)} = \Lambda_0 \Xi, \text{ sur } \Sigma \text{ et } T^{(2)} \rightarrow 0, \text{ à l'infini.}$$

Puis $\rho^{(2)} = p^{(2)} - T^{(2)}$ et ensuite $v_k^{(2)}$ et $p^{(4)}$ se détermine à partir des deux équations (3.7c) et (3.7d) auxquelles il faut associer les conditions:

$$v_k^{(2)} = 0, \text{ sur } \Sigma; \quad v_k^{(2)} \text{ et } p^{(4)} \rightarrow 0, \text{ à l'infini.} \quad (3.8)$$

$$3.4 > \omega > 2$$

Ce cas est intéressant en ce sens que l'on ne peut plus supposer que la relation (3.2) a lieu (!). Il faut, pour obtenir une dégénérescence significative, que $\delta = 2$ et on obtient pour $T^{(2)}$ l'équation linéaire non homogène (3.7b) qui doit être alors résolue avec des conditions nulles:

$$T^{(2)} = 0, \text{ sur } \Sigma \text{ et à l'infini.} \quad (3.9)$$

Pour récupérer la condition à la limite pour la température sur Σ il faut faire intervenir un terme de la forme $M_\infty^{-\omega} T^{(\omega)}$ dans le développement asymptotique de la température; la fonction $T^{(\omega)}$ satisfaisant au problème (3.3c), (3.5).

Ainsi, dans ce cas on doit admettre les développements asymptotique suivant

$$\begin{aligned} v_k &= v_k^{(0)} + M_\infty^{-2} v_k^{(2)} + M_\infty^{-\omega} v_k^{(\omega)} + \dots, \\ p &= 1 + M_\infty^{-2} p^{(2)} + M_\infty^{-4} p^{(4)} + M_\infty^{-2+\omega} p^{(2+\omega)} + \dots, \\ \rho &= 1 + M_\infty^{-2} \rho^{(2)} + M_\infty^{-\omega} \rho^{(\omega)} + \dots, \quad T = 1 + M_\infty^{-2} T^{(2)} + M_\infty^{-\omega} T^{(\omega)} + \dots \end{aligned}$$

Les fonctions $v_k^{(0)}$ et $p^{(2)}$ satisfont toujours à (2.3) et (2.4). Quant aux fonctions $v_k^{(2)}$, $p^{(4)}$, $\rho^{(2)}$ et $T^{(2)}$ elles satisfont aux équations (3.7) auxquelles il faut associer les conditions (3.8) et (3.9). Enfin, la fonction $T^{(\omega)}$ satisfait, comme nous l'avons déjà noté, au problème (3.3c), (3.5), puis $\rho^{(\omega)} = -T^{(\omega)}$ et $v_k^{(\omega)}$ et $p^{(2+\omega)}$ au problème (3.3a), (3.4), (3.6).

En conclusion, notons que l'on peut aisément étendre cette méthode au cas où la condition pour la température sur la surface Σ^* est de la forme:

$$-k_0^* \frac{\partial T^*}{\partial x_k^*} n_k^* = \Delta \phi_0^* \Phi, \text{ sur } \Sigma^*,$$

où $\Delta\phi_0^*$ est un flux caractéristique constant lié à Φ , supposé donné et $\mathbf{n}^* = \{n_h^*\}$ le vecteur unité de la normale extérieure à Σ^* .

Nous profitons de l'occasion pour remercier le Professeur J. P. Guiraud de l'Université de Paris VI pour les discussions et suggestions concernant, plus particulièrement, le cas de $\omega > 2$.

Notons, enfin, que la seconde partie de ce travail sera consacrée au cas des fluides pesants en rotation, ce qui nous permettra, en particulier, d'obtenir de manière rationnelle les équations dites de Boussinesq et les conditions aux limites qui doivent leur être associées.

REFERENCES

1. O. Janzen. Beitrag zu einer Theorie der stationären Stromung kompressibler Flüssigkeiten. Phys. Z., 1913, 14, 639.
2. J. Rayleigh. On the flow of compressible fluid past an obstacle. Philos. Mag., 1916, 32, 1-6.
3. H. Viviand. Etude des écoulements instationnaires à faible nombre de Mach avec application au bruit aérodynamique. J. Mécanique, 1970, 9, No. 4, 573-599.
4. S. C. Crow. Aérodynamical sound emission as a singular perturbation problem. Studies Appl. Math., 1970, 49, No. 1, 21-44.
5. M. Van Dyke. Perturbation methods in fluid mechanics. New York, Acad. Press, 1964.
6. P. A. Lagerström. Laminar flow theory in theory of laminar flows. In High Speed aerodynamics and jet propulsion. Vol. IV. Sect. B. Princeton, Princeton Univ. Press, 1964, 20-285.
7. W. Eckhaus. Matched asymptotic expansions and singular perturbations. Amsterdam, North-Holland, 1973.

A DIFFERENCE SCHEME FOR THE PROBLEM OF THE STRONG BENDING OF THIN PLATES*

C. N. VOLOSHANOVSKAYA and M. M. KARCHEVSKII

Kazan'

(Received 23 April 1975; revised 4 November 1975)

THE FIRST boundary value problem for a system of non-linear differential equations describing the strong bending of flexible thin plates is discussed.

In this paper we construct and investigate a difference scheme for a system of non-linear differential equations describing the strong bending of a thin flexible plate with displacements occurring under the action of an arbitrarily directed external load (see, for example, [1]).

Sufficient conditions for the existence and uniqueness of the solution of the differential boundary value problem are first established.

Questions of the existence of the solution in the case of a normal load have previously been considered by many authors (see, for example, [2, 3]).

The method of constructing the difference scheme used in this paper is based on the approximation of the integral identity (variation equation) by an accumulator [4-6]. Then the fundamental properties of the differential problem are preserved for the difference scheme, which permits existence and uniqueness conditions of the solution of the difference scheme to be obtained

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 183-195, 1977.

comparatively simply.

It is shown that in conditions of uniqueness the difference scheme has accuracy $O(h^2)$ on a fairly smooth solution.

1. Statement of the problem. Investigation of the existence and uniqueness of the solution

It is known that the strong bending of thin flexible plates can be described by the following system of equations [1]:

$$\begin{aligned} \frac{\partial}{\partial x_1}(\varepsilon_1 + \nu \varepsilon_2) + \frac{1-\nu}{2} \frac{\partial \gamma}{\partial x_2} &= -P_1, \\ \frac{\partial}{\partial x_2}(\varepsilon_2 + \nu \varepsilon_1) + \frac{1-\nu}{2} \frac{\partial \gamma}{\partial x_1} &= -P_2, \\ \frac{\alpha^4}{12} \Delta^2 w - \alpha^2 \left\{ \frac{\partial}{\partial x_1} \left[(\varepsilon_1 + \nu \varepsilon_2) \frac{\partial w}{\partial x_1} + \frac{1-\nu}{2} \gamma \frac{\partial w}{\partial x_2} \right] \right. \\ &\quad \left. + \frac{\partial}{\partial x_2} \left[(\varepsilon_2 + \nu \varepsilon_1) \frac{\partial w}{\partial x_2} + \frac{1-\nu}{2} \gamma \frac{\partial w}{\partial x_1} \right] \right\} = -Q, \end{aligned} \quad (1)$$

where $\alpha = h'(\text{mes}^{1/2} \Omega')^{-1}$, h' is the thickness of the plate, Ω' is the region occupied by the plate, $\varepsilon_1, \varepsilon_2, \gamma$ are deformations of elongation and shear of the mean surface,

$$\varepsilon_i = \frac{\partial u_i}{\partial x_i} + \frac{\alpha^2}{2} \left(\frac{\partial w}{\partial x_i} \right)^2, \quad i=1, 2, \quad \gamma = \frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} + \alpha^2 \frac{\partial w}{\partial x_1} \frac{\partial w}{\partial x_2};$$

x_1, x_2 are dimensions coordinates, $u(u_1, u_2)$ is the dimensionless vector of the displacements in the $x_1 x_2$ -plane, w is the vertical displacement of points of the mean surface, and $P(P_1, P_2), Q$ are the components of the external load.

The dimensionless variables are introduced as follows:

$$\begin{aligned} x_i &= x_i' (\text{mes}^{1/2} \Omega')^{-1}, & u_i &= u_i' (\text{mes}^{1/2} \Omega')^{-1}, & w &= w'/h', \\ P_i &= P_i' \text{mes}^{1/2} \Omega'/B, & Q &= Q'h'/B. \end{aligned}$$

Here $B = Eh'/(1-\nu^2)$ is the longitudinal rigidity, E is Young's modulus, and $0 < \nu < 1$ is Poisson's ratio.

We will consider the case of a plate rigidly clamped along the contour:

$$u|_{\Gamma} = 0, \quad w|_{\Gamma} = 0, \quad \frac{\partial w}{\partial n} \Big|_{\Gamma} = 0, \quad (2)$$

Γ — is the boundary of the domain Ω , and n is the normal to Γ .

We define the generalized solution of problem (1), (2) as a vector function (u, w) , $u_1, u_2 \in \dot{W}_2^{(1)}$, $w \in \dot{W}_2^{(2)}$, satisfying the integral identity.

$$\begin{aligned} & \int_{\Omega} \left[(\varepsilon_1 + \nu \varepsilon_2) \left(\frac{\partial \eta_1}{\partial x_1} + \alpha^2 \frac{\partial w}{\partial x_1} \frac{\partial \xi}{\partial x_1} \right) + (\varepsilon_2 + \nu \varepsilon_1) \left(\frac{\partial \eta_2}{\partial x_2} + \alpha^2 \frac{\partial w}{\partial x_2} \frac{\partial \xi}{\partial x_2} \right) \right. \\ & + \frac{1-\nu}{2} \gamma \left(\frac{\partial \eta_1}{\partial x_2} + \frac{\partial \eta_2}{\partial x_1} + \alpha^2 \frac{\partial w}{\partial x_2} \frac{\partial \xi}{\partial x_1} + \alpha^2 \frac{\partial w}{\partial x_1} \frac{\partial \xi}{\partial x_2} \right) \\ & \left. + \frac{\alpha^4}{12} \Delta w \Delta \xi \right] dx = \int_{\Omega} (P_1 \eta_1 + P_2 \eta_2 + Q \xi) dx \end{aligned} \quad (3)$$

for any $\eta_i \in \dot{W}_2^{(1)}$, $\xi \in \dot{W}_2^{(2)}$.

In this paper the following *a priori* estimates of the solution of problem (1), (2) are essentially used.

Lemma 1

For any $P_1, P_2 \in W_2^{(-1)}$ the *a priori* estimate

$$\|u\|_{W_2^{(1)}} \leq \frac{c^2 \alpha^2}{2} \left(\frac{3-\nu}{1-\nu} \right)^{1/2} \|w\|_{W_2^{(2)}}^2 + \frac{2}{1-\nu} \|P\|_{W_2^{(-1)}}; \quad (4)$$

holds; if the condition

$$\|P\|_{W_2^{(-1)}} \leq \frac{1}{c^2 \alpha^2} \left(\frac{1-\nu}{3-\nu} \right)^{1/2} \left(\frac{\alpha^4}{12} - \delta \right), \quad \delta > 0, \quad (5)$$

is also satisfied, then

$$\|w\|_{W_2^{(2)}}^2 \leq K(\delta) (\|P\|_{W_2^{(-1)}}^2 + \|Q\|_{W_2^{(2)}}^2), \quad (6)$$

where

$$K(\delta) = \max \left(\frac{8}{\delta(1-\nu)}, \frac{1}{\delta^2} \right),$$

$$\|u\|_{W_2^{(1)}}^2 = \|u_1\|_{W_2^{(1)}}^2 + \|u_2\|_{W_2^{(1)}}^2, \quad \|P\|_{W_2^{(-1)}}^2 = \|P_1\|_{W_2^{(-1)}}^2 + \|P_2\|_{W_2^{(-1)}}^2,$$

c is the constant from the inequality

$$\|v\|_{\dot{W}_2^{(1)}} \leq c \|v\|_{\dot{W}_2^{(2)}}, \quad (7)$$

valid for any function v of $\dot{W}_2^{(2)}$ [7].

Proof. 1. Putting in (3) $\eta_1 = u_1$, $\eta_2 = u_2$, $\xi = 0$ and using the Cauchy inequality and the inequality (7), we obtain

$$I = \int_{\Omega} \left[\left(\frac{\partial u_1}{\partial x_1} \right)^2 + 2\nu \frac{\partial u_1}{\partial x_1} \frac{\partial u_2}{\partial x_2} + \left(\frac{\partial u_2}{\partial x_2} \right)^2 + \frac{1-\nu}{2} \left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} \right)^2 \right] dx$$

$$\begin{aligned}
&= -\frac{\alpha^2}{2} \int_{\Omega} \left[\left(\frac{\partial w}{\partial x_1} \right)^2 \left(\frac{\partial u_1}{\partial x_1} + \nu \frac{\partial u_2}{\partial x_2} \right) + \left(\frac{\partial w}{\partial x_2} \right)^2 \left(\frac{\partial u_2}{\partial x_2} + \nu \frac{\partial u_1}{\partial x_1} \right) \right. \\
&\quad \left. + (1-\nu) \frac{\partial w}{\partial x_1} \frac{\partial w}{\partial x_2} \left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} \right) \right] dx + \int_{\Omega} (P_1 u_1 + P_2 u_2) dx \\
&\leq \|P\|_{W_2^{(-1)}} \|u\|_{W_2^{(1)}} + \frac{\alpha^2}{2} \left(\frac{3-\nu}{2} \right)^{1/2} c^2 \|w\|_{W_2^{(2)}}^2 I^{1/2}.
\end{aligned} \quad (\text{cont'd})$$

Then using the easily proved inequality

$$\|u\|_{W_2^{(1)}} \leq \left(\frac{2}{1-\nu} \right)^{1/2} I^{1/2},$$

we obtain (4).

2. Now putting $\eta_1 = 2u_1$, $\eta_2 = 2u_2$, $\xi = w$, in (3) we obtain

$$\frac{\alpha^4}{12} \|w\|_{W_2^{(2)}}^2 \leq 2 \|P\|_{W_2^{(-1)}} \|u\|_{W_2^{(1)}} + \|Q\|_{W_2^{(-2)}} \|w\|_{W_2^{(2)}}.$$

Using (4), we have

$$\begin{aligned}
\frac{\alpha^4}{12} \|w\|_{W_2^{(2)}}^2 &\leq 2 \|P\|_{W_2^{(-1)}} \left(\frac{c^2 \alpha^2}{2} \left(\frac{3-\nu}{1-\nu} \right)^{1/2} \|w\|_{W_2^{(2)}} + \frac{2}{1-\nu} \|P\|_{W_2^{(-1)}} \right) \\
&\quad + \|Q\|_{W_2^{(-2)}} \|w\|_{W_2^{(2)}},
\end{aligned}$$

or

$$\begin{aligned}
\left(\frac{\alpha^4}{12} - c^2 \alpha^2 \left(\frac{3-\nu}{1-\nu} \right)^{1/2} \|P\|_{W_2^{(-1)}} \right) \|w\|_{W_2^{(2)}}^2 &\leq \frac{4}{1-\nu} \|P\|_{W_2^{(-1)}}^2 \\
&\quad + \frac{1}{2\delta_1} \|Q\|_{W_2^{(-2)}}^2 + \frac{\delta_1}{2} \|w\|_{W_2^{(2)}}^2.
\end{aligned}$$

Choosing $\delta_1 = \delta$, we obtain

$$\|w\|_{W_2^{(2)}}^2 \leq \frac{8}{\delta(1-\nu)} \|P\|_{W_2^{(-1)}}^2 + \frac{1}{\delta^2} \|Q\|_{W_2^{(-2)}}^2.$$

Using this lemma and the method of [3], we can prove the following theorem.

Theorem 1

Let $P_1, P_2 \in W_2^{(-1)}$, $Q \in W_2^{(-2)}$ and condition (5) be satisfied, then problem (1), (2) has at least one generalized solution.

To investigate the uniqueness of the solution of problem (1), (2) we will require Lemma 2.

Lemma 2

Let condition (5) be satisfied, then

$$\int_{\Omega} \left(\varepsilon_1^2 + 2\nu \varepsilon_1 \varepsilon_2 + \varepsilon_2^2 + \frac{1-\nu}{2} \gamma^2 \right) dx \leq \frac{2}{1-\nu} \|P\|_{W_2^{(-1)}}^2 + \frac{1}{8\delta} \|Q\|_{W_2^{(-2)}}^2. \quad (8)$$

Proof. Putting $\eta_1=2u_1$, $\eta_2=2u_2$, $\xi=w$, in (3), we obtain

$$\int_{\Omega} \left\{ \frac{\alpha^4}{12} (\Delta w)^2 + 2 \left[\varepsilon_1^2 + 2\nu \varepsilon_1 \varepsilon_2 + \varepsilon_2^2 + \frac{1-\nu}{2} \gamma^2 \right] \right\} dx \\ = 2 \int_{\Omega} (P_1 u_1 + P_2 u_2) dx + \int_{\Omega} Q w dx \leq 2 \|P\|_{W_2^{(-1)}} \|u\|_{W_2^{(1)}} + \|Q\|_{W_2^{(-1)}} \|w\|_{W_2^{(1)}}.$$

Using (4), we have

$$\frac{\alpha^4}{12} \int_{\Omega} (\Delta w)^2 dx + 2 \int_{\Omega} \left(\varepsilon_1^2 + 2\nu \varepsilon_1 \varepsilon_2 + \varepsilon_2^2 + \frac{1-\nu}{2} \gamma^2 \right) dx \\ \leq 2 \|P\|_{W_2^{(-1)}} \left(\frac{c^2 \alpha^2}{2} \left(\frac{3-\nu}{1-\nu} \right)^{1/2} \|w\|_{W_2^{(2)}}^2 + \frac{2}{1-\nu} \|P\|_{W_2^{(-1)}} \right) \\ + \frac{1}{2\delta_1} \|Q\|_{W_2^{(-1)}}^2 + \frac{\delta_1}{2} \|w\|_{W_2^{(2)}}^2.$$

Putting $\delta_1/2=\delta$, in this, we obtain (8).

Lemma 3

Let (u, w) , $(u^{(1)}, w^{(1)})$ be generalized solutions of problem (1), (2), and let

$$\left[\int_{\Omega} \left(\varepsilon_1^2 + 2\nu \varepsilon_1 \varepsilon_2 + \varepsilon_2^2 + \frac{1-\nu}{2} \bar{\gamma}^2 \right) dx \right]^{1/2} \leq \frac{\alpha^2}{12c^2} \left(\frac{2}{3-\nu} \right)^{1/2}, \quad (9)$$

where $\varepsilon_i = (\varepsilon_i + \varepsilon_i^{(1)})/2$, $\bar{\gamma} = (\gamma + \gamma^{(1)})/2$. Then $u=u^{(1)}$, $w=w^{(1)}$.

Proof. Putting $\eta_i = u_i - u_i^{(1)}$, $i=1, 2$, $\xi = w - w^{(1)}$ in the integral identities for (u, w) ($u^{(1)}, w^{(1)}$) and subtracting these identities term by term, after simple but laborious transformations we obtain

$$\int_{\Omega} \left[(\varepsilon_1 - \varepsilon_1^{(1)})^2 + 2\nu (\varepsilon_1 - \varepsilon_1^{(1)}) (\varepsilon_2 - \varepsilon_2^{(1)}) + (\varepsilon_2 - \varepsilon_2^{(1)})^2 \right. \\ \left. + \frac{1-\nu}{2} (\gamma - \gamma^{(1)})^2 \right] dx + \frac{\alpha^4}{12} \int_{\Omega} (\Delta (w - w^{(1)}))^2 dx + I = 0,$$

where

$$I = \frac{\alpha^2}{2} \int_{\Omega} \left[\left(\frac{\partial (w - w^{(1)})}{\partial x_1} \right)^2 (\varepsilon_1 + \nu \varepsilon_2) \right. \\ \left. + \left(\frac{\partial (w - w^{(1)})}{\partial x_2} \right)^2 (\varepsilon_2 + \nu \varepsilon_1) + (1-\nu) \bar{\gamma} \frac{\partial (w - w^{(1)})}{\partial x_1} \frac{\partial (w - w^{(1)})}{\partial x_2} \right] dx.$$

It is easy to see that

$$\begin{aligned}
|I| &\leq \frac{\alpha^2}{2} \left\{ \int_{\Omega} \left[\left(\frac{\partial(w-w^{(1)})}{\partial x_1} \right)^4 + (1-\nu) \left(\frac{\partial(w-w^{(1)})}{\partial x_1} \right)^2 \right. \right. \\
&\quad \times \left. \left. \left(\frac{\partial(w-w^{(1)})}{\partial x_2} \right)^2 + \left(\frac{\partial(w-w^{(1)})}{\partial x_2} \right)^4 \right] dx \right\}^{1/2} \\
&\quad \times \left\{ \int_{\Omega} [(\varepsilon_1 + \nu \varepsilon_2)^2 + (1-\nu) \bar{\gamma}^2 + (\varepsilon_2 + \nu \varepsilon_1)^2] dx \right\}^{1/2}.
\end{aligned}$$

From this, by (9), we have

$$\begin{aligned}
&\int_{\Omega} \left[(\varepsilon_1 - \varepsilon_1^{(1)})^2 + 2\nu(\varepsilon_1 - \varepsilon_1^{(1)})(\varepsilon_2 - \varepsilon_2^{(1)}) + (\varepsilon_2 - \varepsilon_2^{(1)})^2 \right. \\
&\quad \left. + \frac{1-\nu}{2} (\gamma - \gamma^{(1)})^2 \right] dx + \|w - w^{(1)}\|_{W_2^{(2)}}^2 \leq 0,
\end{aligned}$$

consequently $u = u^{(1)}$, $w = w^{(1)}$. The lemma is proved.

Remark 1. Lemma 3 shows that if the potential energy of deformation of the mean surface is sufficiently small, the plate has a unique equilibrium shape.

Theorem 2 follows directly from Lemmas 2, 3.

Theorem 2

Let condition (5) be satisfied, and let

$$\left(\frac{1}{8\delta} \|Q\|_{W_2^{(-2)}} + \frac{2}{1-\nu} \|P\|_{W_2^{(-1)}} \right)^{1/2} \leq \frac{\alpha^2}{12c^2} \left(\frac{2}{3-\nu} \right)^{1/2}. \quad (10)$$

Then problem (1), (2) has a unique generalized solution.

Remark 2. Let $P = 0$. Then δ can be put equal to $\alpha^4/12$ and condition (10) assumes the form

$$\|Q\|_{W_2^{(-2)}} \leq \frac{\alpha^4}{6c^2} [3(3-\nu)]^{-1/2}.$$

Then, by (6), for a deflection w we obtain the inequality

$$\max_x |w| \leq \|w\|_{W_2^{(2)}} \leq \frac{12}{\alpha^4} \|Q\|_{W_2^{(-2)}} \leq \frac{2}{c^2 [3(3-\nu)]^{1/2}}.$$

For a dimensional deflection w' we obtain

$$\max_x |w'| \leq \frac{2h'}{c^2 [3(3-\nu)]^{1/2}}.$$

In particular, when the plate is rectangular we have $c^2 = 1/\pi$, consequently,

$$\max_x |w'| \leq \frac{2\pi}{[3(3-\nu)]^{1/2}} h' \leq 2.5h'.$$

This estimate gives an idea of the amount of deflection permitted by the conditions of uniqueness of the solution.

2. Construction of the difference scheme

In what follows we will confine ourselves to the case of a rectangular plate

$$\Omega = \{x | 0 \leq x_1, x_2 \leq 1\}.$$

We construct on Ω a mesh with steps h_1, h_2 along the x_1, x_2 axes.

We introduce notations for the subsets of the mesh, the difference ratios and the sums of the mesh functions:

$$\begin{aligned}\bar{\omega} &= \{x | x = (i_1 h_1, i_2 h_2), i_k = -1, 0, \dots, N_k + 1, N_k h_k = 1\}, \\ \omega &= \{x | x = (i_1 h_1, i_2 h_2), i_k = 1, 2, \dots, N_k - 1\}, \\ \gamma_i &= \{x | x \in \bar{\omega}, x_i = 0 \text{ or } x_i = 1, 0 \leq x_j \leq 1, i \neq j\}, \\ \bar{\gamma}_i &= \{x | x \in \bar{\omega}, x_i = -h_i \text{ or } x_i = 1 + h_i, -h_j \leq x_j \leq 1 + h_j, i \neq j\}, \\ \gamma &= \gamma_1 + \gamma_2, \quad \bar{\gamma} = \bar{\gamma}_1 + \bar{\gamma}_2.\end{aligned}$$

Let $r = (r_1, r_2)$ be a vector whose coordinates can assume the values ± 1 . We put

$$\begin{aligned}\varepsilon_{ir}(y, v) &\equiv \varepsilon_{ir} = \partial_{r_i} y_i + \frac{\alpha^2}{2} (\partial_{r_i} v)^2, \\ \gamma_r(y, v) &\equiv \gamma_r = \partial_{r_1} y_1 + \partial_{r_2} y_2 + \alpha^2 \partial_{r_1} v \partial_{r_2} v.\end{aligned}$$

Here

$$\partial_{r_i} y = \begin{cases} y_{x_i}, & r_i = +1, \\ y_{\bar{x}_i}, & r_i = -1, \end{cases} \quad (y, z)_r = ((y, z)_{r_1}, 1)_{r_2},$$

where

$$\begin{aligned}(y, z)_{r_i} &= \begin{cases} h_i \sum_{j_i=1}^{N_i} y_{j_i j_2} z_{j_i j_2}, & r_i = +1, \\ h_i \sum_{j_i=0}^{N_i-1} y_{j_i j_2} z_{j_i j_2}, & r_i = -1, \end{cases} \\ (y, z) &= h_1 h_2 \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2-1} y_{i_1 i_2} z_{i_1 i_2}, \quad [y, z] = \frac{1}{4} \sum_r (y, z)_r, \\ \|y\|_{1,p} &= \frac{1}{4} \sum_{i=1}^2 \sum_r ((\partial_{-r_i} y)^p, 1)_r, \quad p > 1, \quad \|y\|_1 = \|y\|_{1,2}, \\ \|y\|_2^2 &= \frac{1}{4} \sum_r ((\Delta y)^2, 1)_r,\end{aligned}$$

$\Delta y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2}$ is Laplace's difference operator;

$$\|y\|_{-k} = \sup_{z \neq 0} \frac{|(y, z)|}{\|z\|_k}, \quad k = 1, 2.$$

We denote by H the linear space of mesh vector-functions $Y(y, v) = (y_1, y_2, v)$, defined on $\bar{\omega}$ and satisfying the boundary condition

$$y|_{\gamma+\tilde{\gamma}} = 0, \quad v|_{\gamma} = 0, \quad v_{x_i}^0|_{\gamma_i} = 0, \quad i = 1, 2. \quad (11)$$

In the construction of a difference scheme for problem (1), (2) we will start from the integral identity (3).

Definition. We say that the mesh vector function $Y(y, v) \in H$ is a solution of the difference scheme for problem (1), (2), if for any mesh function $\chi(\eta, \xi) \in H$ the summation identity

$$\begin{aligned} A(Y, \chi) = & \frac{1}{4} \sum_r \left((\varepsilon_{1r} + v\varepsilon_{2r}) (\partial_r \eta_1 + \alpha^2 \partial_r v \partial_r \xi) \right. \\ & + (\varepsilon_{2r} + v\varepsilon_{1r}) (\partial_r \eta_2 + \alpha^2 \partial_r v \partial_r \xi) + \frac{1-v}{2} \gamma_r (\partial_r \eta_2 + \partial_r \eta_1 \\ & \left. + \alpha^2 \partial_r v \partial_r \xi + \alpha^2 \partial_r v \partial_r \xi) + \frac{\alpha^4}{12} \Delta v \Delta \xi, 1 \right) = (\varphi_1, \eta_1) + (\varphi_2, \eta_2) + (\theta, \xi), \end{aligned} \quad (12)$$

is satisfied, where $\varphi_1, \varphi_2, \theta$ are the mesh functions approximating the functions P_1, P_2, Q respectively:

$$\|P_1 - \varphi_1\|_{-1} = O(h^2), \quad \|Q - \theta\|_{-2} = O(h^2).$$

It is easy to see [4], that in searching for the function Y there is then obtained an equation of the form

$$AY = \Phi, \quad \Phi = (\varphi_1, \varphi_2, \theta). \quad (13)$$

To obtain the explicit form of the first equation of the system (13) at the point $x_0 \in \omega$, it is sufficient to put in (12)

$$(\eta, \xi) = (\delta_h(x_0 - x), 0, 0),$$

where

$$\delta_h(x_0 - x) = \begin{cases} 0, & x \neq x_0, \\ 1/h_1 h_2, & x = x_0, \end{cases}$$

if x_0 is at a distance from $\tilde{\gamma}_i$ greater than $2h_i$; otherwise $\delta_h(x_0 - x)$ is put equal to $1/h_1 h_2$ at points $\tilde{\gamma}$ closest to x_0 . The remaining equations of the system (13) are obtained similarly.

We note that from the difference scheme (13) we obtain as a special case the difference scheme for the plane problem of the theory of elasticity [8] (for $\theta = 0, v = 0$) and that for the problem of small deflections of the plate (for $y = 0, \varphi = 0$).

3. Investigation of the solvability of the difference scheme

In the investigation of the existence and uniqueness of the solution of the difference scheme we require the following auxiliary results.

Lemma 4

For any mesh function v satisfying conditions (11) the following inequality holds:

$$\|v\|_{1,4} \leq c_1 \|v\|_2, \quad c_1 = \text{const.} \quad (14)$$

Proof. Let \tilde{v} be a mesh function defined on $\bar{\omega}$ and equal to zero on $\gamma + \bar{\gamma}$. Then, using the difference analog of the embedding theorem from $W_2^{(1)}$ into L_q [12], we obtain $\|\tilde{v}\|_{1,4} \leq c_1 \|\tilde{v}\|_2$.

Now let $v \in H$; we put $\tilde{v} = v(x)$ for $x \in \omega$, $\tilde{v} = 0$, $x \in \bar{\omega}$. It is easy to see that $\|\tilde{v}\|_{1,4} = \|v\|_{1,4}$, $\|\tilde{v}\|_2 = \|v\|_2$, consequently (14) is satisfied.

Prior estimates similar to (4), (6) hold for the difference scheme (18). More precisely the following lemma holds.

Lemma 5

For any δ_1, δ_2 the estimate

$$\|y\|_1 \leq \frac{c_1^2 \alpha^2}{2} \left(\frac{3-v}{1-v} \right)^{1/2} \|v\|_2^2 + \frac{2}{1-v} \|\varphi\|_{-1}. \quad (15)$$

holds for the solution of problem (13). If the condition

$$\|\varphi\|_{-1} \leq \frac{1}{c_1^2 \alpha^2} \left(\frac{1-v}{3-v} \right)^{1/2} \left(\frac{\alpha^4}{12} - \delta \right), \quad \delta > 0, \quad (16)$$

is also satisfied, then

$$\|v\|_2^2 \leq K(\delta) (\|\varphi\|_{-1}^2 + \|\theta\|_{-2}^2), \quad (17)$$

where $K(\delta) = \max(8/\delta(1-v), 1/\delta^2)$,

$$\begin{aligned} & \frac{1}{4} \sum_r \left(\varepsilon_{1r}^2 + 2v\varepsilon_{1r}\varepsilon_{2r} + \varepsilon_{2r}^2 + \frac{1-v}{2} \gamma_r^2, 1 \right) \\ & \leq \frac{2}{1-v} \|\varphi\|_{-1}^2 + \frac{1}{8\delta} \|\theta\|_{-2}^2. \end{aligned} \quad (18)$$

The proof of inequalities (15), (17), (18) is very similar to the proof of the corresponding inequalities (4), (6), (8).

Theorem 3

Let condition (16) be satisfied. Then the difference scheme (13) has at least one solution for any θ .

Proof. It is easy to see that the difference scheme can be written as a system of two operator equations

$$A_1 y = F(v, \varphi), \quad (19)$$

$$A_2(v, y) = \vartheta, \quad (20)$$

where A_1 is a linear operator. By (15), Eq. (19) is uniquely solvable for y for any v, φ . Therefore, system (19), (20) is equivalent to the equation

$$A_2(v, A_1^{-1}F(v, \varphi)) = \vartheta. \quad (21)$$

We now show that an $R > 0$, exists such that $(A_2(v, A_1^{-1}F(v, \varphi)) - \vartheta, v) \geq 0$ for $\|v\|_2 \geq R$. Then, by a well-known topological lemma (see, for example, [10], p. 66), Eq. (24) will have at least one solution. It is easy to see that

$$\begin{aligned} I &= (A_2(v, A_1^{-1}F(v, \varphi)), v) - (\vartheta, v) = (A_2(v, A_1^{-1}F(v, \varphi)), v) \\ &- (\vartheta, v) + (A_1 y, y) - (F(v, \varphi), y) = \frac{\alpha^4}{12} \|v\|_2^2 \\ &+ \frac{1}{4} \sum_r \left(\left(\varepsilon_{1r}^2 + 2\nu \varepsilon_{1r} \varepsilon_{2r} + \varepsilon_{2r}^2 + \frac{1-\nu}{2} \gamma_r^2 \right), 1 \right)_r - (\varphi_1, y_1) \\ &- (\varphi_2, y_2) - (\vartheta, v). \end{aligned}$$

Now using the estimate (15), we obtain

$$I \geq \frac{\alpha^4}{12} \|v\|_2^2 - 2\|\varphi\|_{-1}\|y\|_1 - \|\vartheta\|_{-2}\|v\|_2 \geq 0$$

for

$$\|v\|_2^2 \geq \frac{2}{\delta} \left(\frac{4}{1-\nu} \|\varphi\|_{-1}^2 + \frac{1}{2\delta} \|\vartheta\|_{-2}^2 \right).$$

Theorem 4

Let condition (16) be satisfied and

$$\left(\frac{1}{8\delta} \|\vartheta\|_{-2}^2 + \frac{2}{1-\nu} \|\varphi\|_{-1}^2 \right)^{1/2} \leq \frac{\alpha^2}{12c_1^2} \left(\frac{2}{3-\nu} \right)^{1/2} - \delta, \quad \delta > 0. \quad (22)$$

Then the difference scheme has a unique solution.

The proof of this theorem is exactly similar to the proof of Theorem 2.

4. Estimation of the rate of convergence of the difference scheme

In investigating the convergence of the solution of the difference scheme (13) to the solution of problem (1), (2) we will suppose that the functions u_1, u_2 are continuously differentiable four times in the domain $\bar{\Omega}$, the function w is six times continuously differentiable in some closed region $\tilde{\Omega} \supset \bar{\Omega}$.

We will require the following auxiliary results.

Lemma 6

Let the vector functions (y, v) , $(y^{(1)}, v^{(1)})$ satisfy the equations

$$A_1 y = F(v, \varphi), \quad (23)$$

$$A_1 y^{(1)} = F(v^{(1)}, \varphi^{(1)}). \quad (24)$$

Then

$$\|y - y^{(1)}\|_1 \leq M [(\|v\|_2 + \|v^{(1)}\|_2) \|v - v^{(1)}\|_2 + \|\varphi - \varphi^{(1)}\|_{-1}]. \quad (25)$$

Here and below we will denote by M , constants independent of the mesh step, possibly different.

Proof. We subtract Eq. (24) term by term from (23) and multiply both sides of the resulting equation scalarly by $y - y^{(1)}$. Then, by the definition of the operator A_1 , we obtain

$$\begin{aligned} & \frac{1}{4} \sum_r \left((\varepsilon_{1r} + v \varepsilon_{2r} - \varepsilon_{1r}^{(1)} - v \varepsilon_{2r}^{(1)}) \partial_{r_1} (y_1 - y_1^{(1)}) \right. \\ & \quad \left. + (\varepsilon_{2r} + v \varepsilon_{1r} - \varepsilon_{2r}^{(1)} - v \varepsilon_{1r}^{(1)}) \partial_{r_2} (y_2 - y_2^{(1)}) \right. \\ & \quad \left. + \frac{1-v}{2} (\gamma_r - \gamma_r^{(1)}) (\partial_{r_2} (y_1 - y_1^{(1)}) + \partial_{r_1} (y_2 - y_2^{(1)})), 1 \right)_r \\ & = (\varphi_1 - \varphi_1^{(1)}, y_1 - y_1^{(1)}) + (\varphi_2 - \varphi_2^{(1)}, y_2 - y_2^{(1)}). \end{aligned}$$

Now collecting on the left side of the last equation the terms containing only $y, y^{(1)}$, and using the Cauchy inequality, we obtain

$$\begin{aligned} & \frac{1}{4} \sum_r \left((\partial_{-r_1} (y_1 - y_1^{(1)}))^2 + 2v \partial_{-r_1} (y_1 - y_1^{(1)}) \partial_{-r_2} (y_2 - y_2^{(1)}) \right. \\ & \quad \left. + (\partial_{-r_2} (y_2 - y_2^{(1)}))^2 + \frac{1-v}{2} (\partial_{-r_2} (y_1 - y_1^{(1)}) + \partial_{-r_1} (y_2 - y_2^{(1)})), 1 \right)_r \\ & \leq M ((\|v\|_2 + \|v^{(1)}\|_2) \|v - v^{(1)}\|_2 + \|\varphi - \varphi^{(1)}\|_{-1}) \|y - y^{(1)}\|_1. \end{aligned} \quad (26)$$

Taking the lower bound of the positive-definite form on the left side of the inequality (26), we obtain (25). The lemma is proved.

Lemma 7

Let

$$Ay = \Phi, \quad Ay^{(1)} = \Phi^{(1)},$$

$$\left[\frac{1}{4} \sum_r \left(\bar{\varepsilon}_{1r}^2 + 2v \bar{\varepsilon}_{1r} \bar{\varepsilon}_{2r} + \bar{\varepsilon}_{2r}^2 + \frac{1-v}{2} \bar{\gamma}_r^2, 1 \right) \right]^{1/2}$$

$$\leq \frac{\alpha^2}{12c_1^2} \left(\frac{2}{3-\nu} \right)^{1/2} - \delta, \quad \delta > 0, \quad (\text{cont'd})$$

then

$$\|w - w^{(1)}\|_2 \leq M (\|\theta - \theta^{(1)}\|_{-2}^2 + \|\varphi - \varphi^{(1)}\|_{-1} \|y - y^{(1)}\|_1). \quad (27)$$

The proof is exactly similar to the proof of Lemma 3.

Lemma 8

Let (u, w) be a solution of problem (1), (2), satisfying the smoothness conditions formulated above. A function $f(x_1, x_2)$ exists such that

$$\begin{aligned} f|_{\gamma_i} &= 0, \quad f_{x_i}^{(0)}|_{\gamma_i} = w_{x_i}^{(0)}|_{\gamma_i}, \quad i = 1, 2, \\ \frac{\partial^\alpha f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}} &= O(h^2), \quad \alpha = \alpha_1 + \alpha_2 \leq 6. \end{aligned} \quad (28)$$

Proof. Following [11, 12], we will construct the function in the form

$$f(x_1, x_2) = f^{(1)}(x_1, x_2) + f^{(2)}(x_1, x_2),$$

where

$$f^{(1)}(x_1, x_2) = a_0(x_2)x_1^3 + a_1(x_2)x_1^2 + a_2(x_2)x_1 + a_3(x_2).$$

We define the coefficients $a_k(x_2)$, $k=0, 1, 2, 3$, from conditions (28) for $i=1$. Then, by the boundary conditions (2), the estimate $d^j a_k / dx_2^j = O(h^2)$, $j=0, 1, \dots, 6$, holds for the coefficients $a_k(x_2)$ and their derivatives. We will seek the function $f^{(2)}(x_1, x_2)$ in the form

$$f^{(2)}(x_1, x_2) = b_0(x_1)x_2^3 + b_1(x_1)x_2^2 + b_2(x_1)x_2 + b_3(x_1),$$

where we determine the coefficients $b_k(x_1)$ from the condition

$$\begin{aligned} f^{(2)}(x_1, x_2) &= w(x_1, x_2) - f^{(1)}(x_1, x_2), \\ f_{x_2}^{(2)}(x_1, x_2) &= w_{x_2}^{(0)}(x_1, x_2) - f_{x_2}^{(1)}(x_1, x_2) \end{aligned} \quad (29)$$

for $(x_1, x_2) \in \gamma_2$.

It is easy to verify that the estimates $d^j b_k / dx_1^j = O(h^2)$, $j=0, 1, \dots, 6$ will then hold for the coefficients.

Therefore, the function $f(x_1, x_2)$ and all its derivatives will be of order $O(h^2)$.

We show that $f(x_1, x_2)$ satisfies the boundary conditions (28).

The boundary conditions for $x_2 = 0, x_2 = 1$ are satisfied by (29). We also note that

$$f^{(2)}|_{\gamma_1} = 0, \quad f_{x_1}^{(2)}|_{\gamma_1} = 0.$$

Indeed, by the conditions imposed on $f^{(1)}$,

$$\begin{aligned} f_{x_2}^{(2)}|_{x_1=h} &= f_{x_2}^{(2)}|_{x_1=-h}, & x_2=0, & x_2=1, \\ f^{(2)}|_{x_1=h} &= f^{(2)}|_{x_1=-h}, & x_2=0, & x_2=1. \end{aligned}$$

Consequently,

$$f^{(2)}(h, x_2) = f^{(2)}(-h, x_2), \quad f^{(2)}(0, x_2) = 0. \quad (30)$$

The corresponding equations for $x_1 = 1$ are verified similarly. The lemma is proved.

Theorem 5

Let conditions (16), (22) be satisfied. Then the solution of the difference scheme (13) converges to the solution of problem (1), (2) at the rate $O(h^2)$:

$$\|y - u\|_1 + \|w - v\|_2 \leq Mh^2, \quad h^2 = h_1^2 + h_2^2.$$

Proof. Using the smoothness conditions for the functions (u, \tilde{w}) , $\tilde{w} = w - f$, as in [4, 5] we can obtain

$$\begin{aligned} A((u, \tilde{w}), (\eta, \xi)) &= (\varphi_1, \eta_1) + (\varphi_2, \eta_2) + (\vartheta, \xi) \\ &+ (\psi_1, \eta_1) + (\psi_2, \eta_2) + (\psi_\theta, \xi) \end{aligned}$$

for any vector function $(\eta, \xi) \in H$.

Here (ψ, ψ_θ) is the approximation error:

$$\|\psi\|_{-1} = O(h^2), \quad \|\psi_\theta\|_{-2} = O(h^2). \quad (31)$$

We now note that by condition (16) a $\delta_0 \geq \delta > 0$, can be found such that for all $h \leq h_0$, $h_0 > 0$

$$\|\varphi + \psi\|_{-1} \leq \frac{\alpha^2}{12c_1^2} \left(\frac{1-\nu}{3-\nu} \right)^{1/2} - \delta_0.$$

Now using Lemma 5, conditions (16) and (22), we obtain

$$\begin{aligned} &\frac{1}{4} \sum_r \left(\varepsilon_{1r}^2(u, \tilde{w}) + 2\nu \varepsilon_{1r}(u, \tilde{w}) \varepsilon_{2r}(u, \tilde{w}) + \varepsilon_{2r}^2(u, \tilde{w}) \right. \\ &\left. + \frac{1-\nu}{2} \gamma_r^2(u, \tilde{w}), 1 \right) \leq \frac{1}{8\delta_0} \|\vartheta + \psi_\theta\|_{-2}^2 + \frac{2}{1-\nu} \|\psi + \varphi\|_{-1}^2 \\ &\leq \left[\frac{\alpha^2}{12c_1^2} \left(\frac{2}{3-\nu} \right)^{1/2} - \delta \right]^2, \quad \delta > 0. \end{aligned}$$

By (18) a similar inequality is satisfied for (v, v) also. Then using Lemma 7, we obtain

$$\|\tilde{w} - v\|_2^2 \leq M(\|\psi_\theta\|_{-2}^2 + \|\psi\|_{-1} \|y - u\|_1),$$

whence, by the inequalities (25), (31), we have

$$\|w-v\|_2 = O(h^2), \quad \|w-v\|_2 \leq \|f\|_2 + \|\tilde{w}-v\|_2 = O(h^2).$$

Applying once more the inequality (25), we finally obtain

$$\|y-u\|_1 = O(h^2).$$

The theorem is proved.

For the numerical realization of the difference scheme (11)–(13) we can use an iterative process of the form

$$A_1 y^{n+1} = F(v^n, \varphi), \quad \bar{\Delta}^2 v^{n+1} = \bar{\Delta}^2 v^n - \tau(A_2(v^n, y^n) - \theta), \quad \tau > 0. \quad (32)$$

The iterative process (32) presupposes at each step the solution of difference schemes for the plane problem of the theory of elasticity and of the biharmonic equation, which can be carried out, for example, by known iterative methods [13].

The convergence of the method (32) for conditions close to (16), (22), can be investigated by the method described in [14, 15].

The authors sincerely thank A. D. Lyashko for suggesting the problem and for his continued interest.

Translated by J. Berry.

REFERENCES

1. VOL'MIR, A. S. *Flexible plates and shells* (Gibkie plastiny i obolochki), Gostekhizdat, Moscow, 1956.
2. VOROVICH, I. I. On the existence of solutions in the non-linear theory of shells. *Izv. Akad. Nauk SSSR. Ser. matem.*, 19, 4, 173–186, 1955.
3. DUBINSKII, Yu. A. On the solvability of the system of equations of the strong bending of plates. *Dokl. Akad. Nauk SSSR*, 175, 5, 1026–1029, 1967.
4. KARCHEVSKII, M. M. and LYASHKO, A. D. Difference schemes for non-linear multidimensional elliptic equations. I. *Izv. vuzov. Matematika*, 1972, No. 11, 23–31; II. 1973, No. 3, 44–52.
5. LYASHKO, A. D. Difference schemes for the problem of the bending of thin plates. In: *Numerical methods of the mechanics of continuous media* (Chisl. metody mekhan. sploshnoi sredy), 4, 1, 71–83, VTs SO Akad. Nauk SSSR, Novosibirsk, 1973.
6. LADYZHENSKAYA, O. A. *Linear and quasilinear equations of elliptic type* (Lineinye i kvazilineinye uravneniya ellipticheskogo tipa), "Nauka", Moscow, 1973.
7. SOBOLEV, S. L. *Some applications of functional analysis in mathematical physics* (Nekotorye primeneniya funktsional'nogo analiza v matematicheskoi fizike), Izd-vo LGU, Leningrad, 1950.
8. BELUKHINA, I. G. Difference schemes for the solution of some statistical problems of the theory of elasticity. *Zh. vychisl. Mat. mat. Fiz.*, 8, 4, 808–823, 1968.
9. SOBOLEV, S. L. On estimates for some sums for functions defined on a mesh. *Izv. akad. Nauk SSSR. Ser. matem.* 4, 5–16, 1940.
10. LIONS, J.-L. *Some methods of solving non-linear boundary value problems* (Nekotorye metody resheniya nelineinykh kraevykh zadach), "Mir", Moscow, 1972.
11. LAPIN, A. V. and LYASHKO, A. D. Investigation of the mesh method for second-order non-linear elliptic equations. *Izv. vuzov. Matematika*, 10, 37–43, 1970.

12. KARCHEVSKII, M. M. On the convergence of the straight line method for fourth-order elliptic equations. *Izv. vuzov. Matematika*, No. 4, 24-27, 1969.
13. SAMARSKII, A. A. *Introduction to the theory of difference schemes* (Vvedenie v teoriyu raznostnykh skhem), "Nauka", Moscow, 1971.
14. D'YAKONOV, E. G. On the solution of some non-linear systems of difference equations. *Dokl. Akad. Nauk SSSR*, 188, 5, 982-985, 1969.
15. KARCHEVSKII, M. M. Iterative schemes for equations with monotonic operators. *Izv. vuzov. Matematika*, No. 5, 32-37, 1971.

NUMERICAL SOLUTION OF THE TWO-DIMENSIONAL PROBLEM OF SHOCK WAVE PROPAGATION IN OUTER SPACE*

L. V. SHIDLOVSKAYA

Moscow

(Received 1 December 1975; revised 24 March 1976)

A METHOD for the numerical solution of the non-stationary two-dimensional problem of the propagation relative to a moving interplanetary medium of a perturbation caused by the emission of finite energy within the limits of a section of a cone, simulating the chromospheric region of the sun occupied by the flare is discussed. The effect of solar gravitation and of the radial component of the magnetic field strength on the motion of the gas is taken into account.

Introduction

The data on the observation of chromospheric solar flares accumulated up to the present time and the measurements of the parameters of the shock waves arising from these flares and propagated in interstellar space, draw attention to topics connected with the propagation of perturbations in the solar wind. The fundamental results of the analytic and numerical investigations of the dynamic processes in the interplanetary medium, contained in [1-6], were obtained from a study of the one-dimensional flow of a plasma on the assumption of spherical symmetry about the sun. However, a continuously increasing amount of experimental data testifies to the fact that the shock wave front close to the earth's orbit does not possess spherical symmetry, and accordingly more and more complex models of the propagation of perturbations in the solar wind have had to be developed. It is known [1, 7] that solar chromospheric flares arise and take place in a comparatively small volume: the area of a flare occupies about 1% of the area of the solar disk, the height of the flare layer is of the order of 10^4 km. The radiation generated by solar flares has a small duration: from 0.5 to 3 hours. Sometimes the radiation is far from radial, since some flares cause geographical effects, although they take place close to the edge of the solar disk. All this suggests that the radiation of particles from the surface of the sun may occur within the limits of some cone, at times possessing a fairly considerable solid angle. The two-dimensional problem of a blast in a moving medium in the hydrodynamic approximation was considered in [8].

In this paper we present a two-dimensional magnetohydrodynamic model of the motion of a plasma, based on a continuous emission of large energies and masses of incandescent gases in the course of 5×10^2 to 5×10^3 sec into an interplanetary medium moving along a radius from the sun. Allowance is made for the interaction of the solar wind with the interplanetary magnetic field, whose source is the total magnetic field of the sun. In view of the extreme complexity of the phenomena

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 196-208, 1977.

considered it is assumed that the azimuthal component of the magnetic field strength vector exerts no significant effect on the motion of the interplanetary gas, although the converse effect of the dynamic processes on the perturbation of the interplanetary magnetic field may be fairly strong. This constraint is permissible since the experimental data confirm the fact that at distances from the sun right up to the earth's orbit the radial component of the magnetic field predominates over the remaining components. Analytic investigations are difficult because of the complex mathematical formulation (a boundary value problem for a system of non-linear equations of a compressed gas with three independent variables). Therefore to obtain a solution of the problem posed numerical calculations were carried out by the modified non-stationary method of "large particles" [9].

1. Statement of the problem

We consider the system of equations of single-fluid hydrodynamics for a completely ionized, non-heat conducting hydrogenous gas using the spherical coordinate system. We suppose that the flow is symmetrical about the polar axis, so that all the quantities are functions of the heliocentric radius r , the polar angle θ and the time t . The polar axis is so chosen that it passes through the centre of the region occupied by the flare, which is a section of a cone (see Fig. 1). The equations include the forces connected with the pressure gradient and solar gravitation, and the effect of dissipative processes is disregarded. We denote by ρ the numerical density, T is the temperature, p is the pressure, u is the velocity of the flow in the radial direction, v is the tangential component in the direction of variation of the angle θ , and $H \equiv H_r$ is the magnetic field strength. The equations of the conservation of mass momentum and energy in Eulerian coordinates have the form

$$\frac{\partial \rho}{\partial t} + \frac{1}{r^2} \frac{\partial}{\partial r} (\rho u r^2) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\rho v \sin \theta) = 0, \quad (1.1)$$

$$\frac{\partial}{\partial t} (\rho u) + \frac{1}{r^2} \frac{\partial}{\partial r} (\rho u^2 r^2) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\rho u v \sin \theta) = -\frac{\partial p}{\partial r} + \rho \frac{v^2}{r} - \rho \frac{GM_s}{r^2} \quad (1.2)$$

$$\begin{aligned} & \frac{\partial}{\partial t} (\rho v) + \frac{1}{r^2} \frac{\partial}{\partial r} (\rho u v r^2) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\rho v^2 \sin \theta) \\ &= -\frac{1}{r} \frac{\partial p}{\partial \theta} - \rho \frac{uv}{r} - \frac{H}{4\pi r} \frac{\partial H}{\partial \theta}, \end{aligned} \quad (1.3)$$

$$\begin{aligned} & \frac{\partial}{\partial t} (\rho E) + \frac{1}{r^2} \frac{\partial}{\partial r} [(\rho E + p) u r^2] \\ &+ \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} [(\rho E + p) v \sin \theta] = -\frac{H v}{4\pi r} \frac{\partial H}{\partial \theta}. \end{aligned} \quad (1.4)$$

Here G is the gravitational constant, M_s is the mass of the sun, E is the specific total energy (not taking into account the energy of the magnetic field):

$$E = 3kT/(m_p + m_e) + (u^2 + v^2)/2 - GM_s/r, \quad (1.5)$$

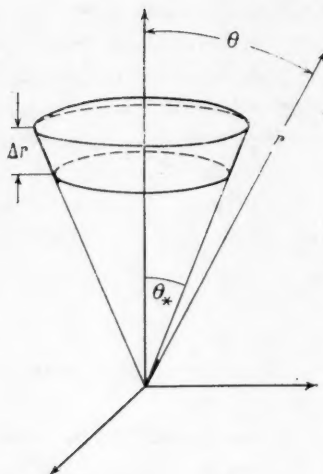


FIG. 1.

m_p and m_e are the masses of a proton and an electron respectively. For the specific internal energy I we have the expression $I = 3kT/(m_p + m_e)$. The equation of state of a perfect gas is written in the form

$$p = 2\rho kT/(m_p + m_e), \quad (1.6)$$

where k is Boltzmann's constant.

To determine the magnetic field strength H in a medium with infinite conductivity from Maxwell's equations, neglecting the displacement current, we have the following equations:

$$\partial H / \partial t + [\partial(Hv \sin \theta) / \partial \theta] / r \sin \theta = 0, \quad (1.7)$$

$$\partial(r^2 H) / \partial r = 0, \quad (1.8)$$

to which in this case the whole system of electrodynamic equations essentially reduces. For the solution to satisfy Eq. (18), it is sufficient to require that it be satisfied by the initial data, therefore Eq. (1.8) will be used only for the construction of the initial parameter distribution of the gas flow. Therefore, system (1.1)–(1.7) is closed.

The electrical field strength E_e and the current density J can be determined from the formulas

$$E_e = -[U \times H], \quad J = \text{rot } H / 4\pi,$$

where U is the gas flow velocity vector.

The general form of system (1.1)–(1.7) can remain the same for dimensional and dimensionless variables, so we will use the latter, choosing as characteristic values of the plasma parameters for r_0 the corresponding experimental data. The motion of the gas occurs in a domain bounded by the spherical surfaces $r = r_0$ and $r = r_N$ with variation of the angle θ from 0 to π .

The initial steady flow of the solar wind is defined as follows. It is assumed that the quiescent solar wind has only a radial velocity component, that is, $v \equiv 0$ at the initial instant. Using this assumption and equating to zero the time derivatives in Eqs. (1.1)–(1.8), we obtain a system of ordinary differential equations whose numerical solution is constructed with the boundary values u_0, ρ_0, p_0 and H_0 specified at the point $r = r_0$. The result of this solution gives the initial distribution of the hydrodynamic parameters and the magnetic field strength.

We assume that at the instant $t = 0$ in the conical section $r_0 \leq r \leq r_0 + \Delta r$, $0 \leq \theta \leq \theta_*$ a sudden change in the parameters of the gas flow has occurred, the new higher values of these parameters $u_{01}, \rho_{01}, p_{01}$ being determined from the conditions on strong discontinuities for a specified initial rate of propagation of the shock wave u_B along the quiescent solar wind [10]. The energy is emitted during the time interval τ . The value of the initial angle θ_* and thickness of the layer Δr in which the perturbation is initially concentrated can be specified arbitrarily, and in performing numerical calculations Δr is taken equal to half the mesh step in the radial direction. Therefore, at the initial instant the distribution of the parameters u, v, ρ, p, H of the quiescent solar wind and the perturbation concentrated in the section of the cone of parameters are specified. It is required to determine the flow of the gas at subsequent instants, until the front of the shock wave reaches the earth's orbit, subject to the following boundary conditions:

$$u, v, \rho, p, H = \begin{cases} u_0, & 0, & \rho_0, & p_0, & H_0 & \text{for } r = r_0, & \theta_* < \theta \leq \pi, & t \geq 0, \\ u_{01}, & 0, & \rho_{01}, & p_{01}, & H_0 & \text{for } r_0 \leq r \leq r_0 + \Delta r, & 0 \leq \theta \leq \theta_*, & t \leq \tau, \\ u_0, & 0, & \rho_0, & p_0, & H_0 & \text{for } r_0 \leq r \leq r_0 + \Delta r, & 0 \leq \theta \leq \theta_*, & t > \tau. \end{cases}$$

The requirement of symmetry of the flow about the polar axis imposes one more boundary condition on the tangential component of the gas velocity: $v = 0$ for $u_{01}, 0, \rho_{01}, p_{01}$. The parameters of the perturbed flow change not only in space, but also in time. To construct the solution of the problem of the non-stationary motion of the gas we use the numerical "large particle" method [9], modified somewhat due to the specific nature of the problem.

2. Method of calculation

We subdivide the domain of integration by a fixed Eulerian mesh into cells whose centres correspond to points with subscripts i, j ; the volume of each cell $V_{ij} = 2\pi (r_i)^2 \sin \theta_j \Delta r \Delta \theta$, where Δr is the step along the radius, $\Delta r = r_{i+1} - r_i$, $\Delta \theta$ is the angular step, and $\Delta \theta = \theta_{j+1} - \theta_j$.

We suppose that at the instant $t = n \Delta t$ for every cell i, j the determined values of the velocity, density, pressure and direction of the magnetic field are known. The problem consists of finding these values at the instant $t = (n+1) \Delta t$.

The calculation of each step is divided into two stages [9]. At the first (Eulerian) stage we neglect the flows of mass through the cell boundaries and determine intermediate values of the

parameters by means of the equations

$$\begin{aligned}\rho \frac{\partial u}{\partial t} &= \rho \frac{v^2}{r} - \frac{\partial p}{\partial r} - \rho \frac{GM_*}{r^2}, \\ \rho \frac{\partial v}{\partial t} &= -\rho \frac{vu}{r} - \frac{1}{r} \frac{\partial p}{\partial \theta} - \frac{H}{4\pi r} \frac{\partial H}{\partial \theta}, \\ \rho \frac{\partial E}{\partial t} &= -\frac{1}{r^2} \frac{\partial}{\partial r} (p u r^2) - \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (p v \sin \theta) - \frac{H v}{4\pi r} \frac{\partial H}{\partial \theta}.\end{aligned}\quad (2.1)$$

The finite-difference equations of first-order accuracy in time and space, corresponding to this system, have the form

$$\begin{aligned}\tilde{u}_{ij}^n &= u_{ij}^n - \frac{\Delta t}{2\rho_{ij}^n r_i^2 \Delta r} [r_{i+1/2}^2 (p_{i+1,j}^n - p_{ij}^n) - r_{i-1/2}^2 (p_{i-1,j}^n - p_{ij}^n)] \\ &\quad - \frac{\Delta t}{r_i^2} [GM_* - r_i (v_{ij}^n)^2],\end{aligned}\quad (2.2)$$

$$\begin{aligned}\tilde{v}_{ij}^n &= v_{ij}^n - \frac{\Delta t}{2\rho_{ij}^n r_i \sin \theta_j \Delta \theta} [\sin \theta_{j+1/2} (p_{i,j+1}^n - p_{ij}^n) \\ &\quad - \sin \theta_{j-1/2} (p_{i,j-1}^n - p_{ij}^n)] - \frac{\Delta t v_{ij}^n u_{ij}^n}{r_i} - \frac{\Delta t H_{ij}^n}{4\pi r_i \rho_{ij}^n \Delta \theta} (H_{i,j+1/2}^n - H_{i,j-1/2}^n),\end{aligned}\quad (2.3)$$

$$\begin{aligned}\tilde{E}_{ij}^n &= E_{ij}^n - \frac{\Delta t}{2\rho_{ij}^n r_i^2 \sin \theta_j \Delta r \Delta \theta} \{ \sin \theta_j \Delta \theta [p_{i+1/2,j}^n r_{i+1/2}^2 (u_{i+1/2,j}^n + \tilde{u}_{i+1/2,j}^n) \\ &\quad - p_{i-1/2,j}^n r_{i-1/2}^2 (u_{i-1/2,j}^n + \tilde{u}_{i-1/2,j}^n)] + r_i \Delta r [p_{i,j+1/2}^n \sin \theta_{j+1/2} (v_{i,j+1/2}^n + \tilde{v}_{i,j+1/2}^n) \\ &\quad - p_{i,j-1/2}^n \sin \theta_{j-1/2} (v_{i,j-1/2}^n + \tilde{v}_{i,j-1/2}^n)] \} - \frac{\Delta t H_{ij}^n v_{ij}^n}{4\pi r_i \rho_{ij}^n \Delta \theta} (H_{i,j+1/2}^n - H_{i,j-1/2}^n).\end{aligned}\quad (2.4)$$

At this stage the difference scheme is stable, as will be shown below on the basis of [9].

We note that another approach to the construction of the difference scheme at the first stage is possible, namely: instead of the equation for the total energy E we can use the equation for the internal energy I . Then, however, the difference scheme will be non-conservative and to improve the stability of calculation at the first stage the pressure p in Eqs. (2.1)–(2.4) must be replaced by the term $(p + q)$, where q is an artificial viscous pressure. If c is the local speed of sound, then the expression for q can be taken in the form [5]

$$\begin{aligned}q_{i+1/2,j}^n &= \begin{cases} \rho_{i+1/2,j}^n (u_{ij}^n - u_{i+1,j}^n) [B c_{i+1/2,j}^n + A^2 (u_{ij}^n - u_{i+1,j}^n)] & \text{for } u_{ij}^n > u_{i+1,j}^n, \\ 0 & \text{for } u_{ij}^n \leq u_{i+1,j}^n, \end{cases} \\ q_{i,j+1/2}^n &= \begin{cases} \rho_{i,j+1/2}^n (v_{ij}^n - v_{i,j+1}^n) [B c_{i,j+1/2}^n + A^2 (v_{ij}^n - v_{i,j+1}^n)] & \text{for } v_{ij}^n > v_{i,j+1}^n, \\ 0 & \text{for } v_{ij}^n \leq v_{i,j+1}^n. \end{cases}\end{aligned}$$

The introduction of the linear term in the expression for q in the case of weak shock waves leads to damping of the oscillations arising behind the shock front. The quadratic term plays a similar role in the consideration of strong shock waves. Experimental measurements of the mean velocities of shock waves at various points of interplanetary space show that in a number of cases

a strong shock wave may experience considerable damping; therefore on the earth's orbit weak shock waves are observed.

At the second (Lagrangian) stage we calculate the density of the mass flow in the motion of the gas through the cell boundaries. Let $M_{i+1/2,j}^n$ be the flow of mass through the boundary $(i + 1/2)$ of the cell i, j in the time Δt , and $M_{i,j+1/2}^n$ the similar flow through the boundary $(j + 1/2)$:

$$M_{i+1/2,j}^n = \begin{cases} 2\pi r_{i+1/2}^2 \sin \theta_j \rho_{ij}^n \tilde{u}_{i+1/2,j}^n \Delta \theta \Delta t & \text{for } \tilde{u}_{i+1/2,j}^n > 0, \\ 2\pi r_{i+1/2}^2 \sin \theta_j \rho_{i+1,j}^n \tilde{u}_{i+1/2,j}^n \Delta \theta \Delta t & \text{for } \tilde{u}_{i+1/2,j}^n < 0, \end{cases}$$

$$M_{i,j+1/2}^n = \begin{cases} 2\pi r_{i+1/2}^2 \sin \theta_j \rho_{ij}^n \tilde{u}_{i,j+1/2}^n \Delta \theta \Delta t & \text{for } \tilde{u}_{i,j+1/2}^n > 0, \\ 2\pi r_i \sin \theta_{j+1/2} \rho_{i,j+1}^n \tilde{v}_{i,j+1/2}^n \Delta r \Delta t & \text{for } \tilde{v}_{i,j+1/2}^n < 0. \end{cases}$$

The third stage of the calculations consists of determining the final values of the flow parameters $\rho, X = (u, v, E)$ corresponding to the instant $t = (n+1)\Delta t$, on the basis of the laws of conservation of mass, momentum and energy for every cell by the formulas

$$\rho_{ij}^{n+1} = \rho_{ij}^n + \frac{M_{i-1/2,j}^n + M_{i,j-1/2}^n - M_{i+1/2,j}^n - M_{i,j+1/2}^n}{V_{ij}}, \quad (2.5)$$

$$X_{ij}^{n+1} = \frac{\rho_{ij}^n X_{ij}^n}{\rho_{ij}^{n+1}} + \frac{\bar{X}_{i-1,j} M_{i-1/2,j}^n + \bar{X}_{i,j-1} M_{i,j-1/2}^n - \bar{X}_{ij} (M_{i+1/2,j}^n + M_{i,j+1/2}^n)}{\rho_{ij}^{n+1} V_{ij}}. \quad (2.6)$$

The specific internal energy and temperature are determined from the relations, true for any instant of time:

$$I_{ij}^{n+1} = E_{ij}^{n+1} - \frac{(u_{ij}^{n+1})^2 + (v_{ij}^{n+1})^2}{2} + \frac{GM_s}{r_i}, \quad T_{ij}^{n+1} = \frac{(m_p + m_e)}{3k} I_{ij}^{n+1}. \quad (2.7)$$

To determine the magnetic field strength at the instant $t = (n+1)\Delta t$ it is proposed to use the following finite-difference equation corresponding to the induction equation (1.7):

$$H_{ij}^{n+1} = H_{ij}^n - \Delta t (H_{ij}^n \tilde{v}_{i,j+1/2}^n \sin \theta_{j+1/2} - H_{i,j-1}^n \tilde{v}_{i,j-1/2}^n \sin \theta_{j-1/2}) / r_i \sin \theta_j \Delta \theta. \quad (2.8)$$

The difference scheme is conservative as a whole, since at each stage the conservation laws are used, and the total variation of mass, momentum and energy after Δt equals the sum of their variations in the first and third stages.

The difference equations in the form (2.5)–(2.8) are true only for the internal cells of the domain considered. We choose the external boundary r_N in such a way that it coincides with the centre of the N -th cell; it is then required simply to determine the flow variables in the fictitious external cell, equal in value to the internal cell adjacent to it. On the boundary r_0 the values of all the hydrodynamic parameters are defined at all instants of time and do not require to be defined.

3. Viscosity effects and the stability condition

The magnetohydrodynamic equations for an inviscid gas are regarded as the fundamental equations for describing the motion of the plasma. However viscous effects arise due to the introduction into the equations of an artificial viscous pressure q and the presence of scheme viscosity, arising from the replacement of the exact differential equations by finite-difference approximations. Topics connected with scheme viscosity are studied in [9]. Expanding the difference operators of the scheme of the problem considered at all stages in Taylor series, we obtain

$$\begin{aligned}
 & \frac{\partial \rho}{\partial t} + \frac{1}{r^2} \frac{\partial}{\partial r} (\rho u r^2) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\rho v \sin \theta) \\
 &= \frac{1}{r^2} \frac{\partial}{\partial r} \left(\epsilon \frac{\partial \rho}{\partial r} \right) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} \left(\eta \frac{\partial \rho}{\partial \theta} \right), \\
 & \frac{\partial}{\partial t} (\rho u) + \frac{1}{r^2} \frac{\partial}{\partial r} (\rho u^2 r^2) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\rho u v \sin \theta) + \frac{\partial p}{\partial r} \\
 &+ \rho \frac{GM_s}{r^2} = \frac{1}{r^2} \frac{\partial}{\partial r} \left[\epsilon \frac{\partial}{\partial r} (\rho u) \right] + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} \left[\eta \frac{\partial}{\partial \theta} (\rho u) \right] - \frac{\partial q}{\partial r}, \\
 & \frac{\partial H}{\partial t} + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (H v \sin \theta) = \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} \left(\eta \frac{\partial H}{\partial \theta} \right), \\
 & \epsilon = |u| r^2 \Delta r / 2, \quad \eta = |v| \sin \theta \Delta \theta / 2.
 \end{aligned} \tag{3.1}$$

Similar investigations can be carried out for Eqs. (1.3)–(1.4) also, producing qualitatively similar results.

Analyzing Eqs. (3.1), we easily notice that even for $q = 0$ terms with ϵ, η are present, which are similar to the dissipative terms of the Navier-Stokes equation; therefore blurring of the shock wave front occurs in the case $q = 0$ also.

As stability conditions for Eqs. (1.1)–(1.7) we use the usual Courant condition (or the speed of sound condition), consisting of the assertion that the time steps must be less than the time interval in which the sound signal reaches the boundaries of the adjacent cells. Taking into account the fact that the perturbation propagates along the solar wind whose velocity is non-zero, and the fact that the perturbed motion takes place along two directions in space, we propose to use the stability condition in the following form:

$$\Delta t = \beta \min [\Delta r / (c + |u|), r \Delta \theta / (c + |v|)], \tag{3.2}$$

where $\beta < 1$ is some coefficient, and $c = (\gamma p / \rho)^{1/2}$ for gases.

We show that with the estimate (3.2) the original system of equations (1.1)–(1.7) is stable in the sense of [9], that is, the sign of the coefficients α_i of the dissipative terms of the differential approximation, containing second-order partial derivatives with respect to the spatial variables, is positive. Expanding the difference equations at all three stages in Taylor series accurate to terms of the first order in time and of the second order in space inclusive, we obtain

$$\frac{\partial \rho}{\partial t} + \frac{1}{r^2} \frac{\partial}{\partial r} (\rho u r^2) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\rho v \sin \theta) = \Delta_1 + \frac{1}{r^2} \left[\frac{\delta r}{2} u r^2 \right]$$

$$\begin{aligned}
& -\frac{(\delta r)^2}{4} \frac{\partial}{\partial r} (ur^2) - \frac{\delta t}{2} r^2 (u^2 + c^2) \left] \frac{\partial^2 \rho}{\partial r^2} + \frac{1}{r \sin \theta} \left[\frac{\delta \theta}{2} v \sin \theta \right. \right. \\
& \left. \left. - \frac{(\delta \theta)^2}{4} \frac{\partial}{\partial \theta} (v \sin \theta) - \frac{\delta t}{2r} (v^2 + c^2) \sin \theta \right] \frac{\partial^2 \rho}{\partial \theta^2} + o(\delta t, \delta r^2, \delta \theta^2), \right. \\
& \frac{\partial}{\partial t} (\rho u) + \frac{1}{r^2} \frac{\partial}{\partial r} (\rho u^2 r^2) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\rho u v \sin \theta) = \Delta_2 \\
& + \frac{1}{r^2} \left[\frac{\delta r}{2} \rho u r^2 - \frac{(\delta r)^2}{2} \frac{\partial}{\partial r} (\rho u r^2) - \frac{(\delta r)^2}{4} u r^2 \frac{\partial \rho}{\partial r} - \frac{\delta t}{2} \rho r^2 \right. \\
& \times (u^2 + v^2) \left] \frac{\partial^2 u}{\partial r^2} + \frac{1}{r \sin \theta} \left[\frac{\delta \theta}{2} \rho v \sin \theta - \frac{(\delta \theta)^2}{4} \frac{\partial}{\partial \theta} (\rho v \sin \theta) \right. \\
& \left. - \frac{(\delta \theta)^2}{4} v \sin \theta \frac{\partial \rho}{\partial \theta} - \frac{\delta t}{2r} \rho (v^2 + c^2) \sin \theta \right] \frac{\partial^2 u}{\partial \theta^2} + o(\delta t, \delta r^2, \delta \theta^2), \\
& \frac{\partial}{\partial t} (\rho v) + \frac{1}{r^2} \frac{\partial}{\partial r} (\rho v u r^2) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\rho v^2 \sin \theta) = \Delta_3 \\
& + \frac{1}{r^2} \left[\frac{\delta r}{2} \rho u r^2 - \frac{(\delta r)^2}{4} \frac{\partial}{\partial r} (\rho u r^2) - \frac{(\delta r)^2}{4} u r^2 \frac{\partial \rho}{\partial r} - \frac{\delta t}{2} \rho r^2 \right. \\
& \times (u^2 + v^2) \left] \frac{\partial^2 v}{\partial r^2} + \frac{1}{r \sin \theta} \left[\frac{\delta \theta}{2} \rho v \sin \theta - \frac{(\delta \theta)^2}{2} \frac{\partial}{\partial \theta} (\rho v \sin \theta) \right. \\
& \left. - \frac{(\delta \theta)^2}{4} v \sin \theta \frac{\partial \rho}{\partial \theta} - \frac{\delta t}{2r} \rho (v^2 + c^2) \sin \theta \right] \frac{\partial^2 v}{\partial \theta^2} + o(\delta t, \delta r^2, \delta \theta^2), \\
& \frac{\partial}{\partial t} (\rho E) + \frac{1}{r^2} \frac{\partial}{\partial r} (\rho E u r^2) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\rho E v \sin \theta) = \Delta_4 \\
& + \frac{1}{r^2} \left[\frac{\delta r}{2} \rho u r^2 - \frac{(\delta r)^2}{4} \rho r^2 \frac{\partial u}{\partial r} - \frac{(\delta r)^2}{4} u r^2 \frac{\partial \rho}{\partial r} \right] \frac{\partial^2 E}{\partial r^2} \\
& + \frac{1}{r \sin \theta} \left[\frac{\delta \theta}{2} \rho v \sin \theta - \frac{(\delta \theta)^2}{4} \rho \sin \theta \frac{\partial v}{\partial \theta} - \frac{(\delta \theta)^2}{4} v \sin \theta \frac{\partial \rho}{\partial \theta} \right] \frac{\partial^2 E}{\partial \theta^2} \\
& + o(\delta t, \delta r^2, \delta \theta^2), \\
& \frac{\partial H}{\partial t} + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (H v \sin \theta) = \Delta_5 + \frac{1}{r \sin \theta} \left[\frac{\delta \theta}{2} H \sin \theta \right. \\
& \left. - \frac{(\delta \theta)^2}{4} \frac{\partial}{\partial \theta} (H \sin \theta) - \frac{\delta t}{2r} (v^2 + c^2) \sin \theta \right] \frac{\partial^2 H}{\partial \theta^2} + o(\delta t, \delta \theta^2),
\end{aligned}$$

where $\Delta_1, \Delta_2, \Delta_3, \Delta_4, \Delta_5$ are terms of the first differential approximation proportional to $\Delta r, r\Delta\theta$ and containing first derivatives. For the calculations we chose

$$\delta r = 0.01, \quad \delta \theta = 0.1, \quad \delta t = 0.0008, \quad 0.1 \leq r \leq 1.1.$$

To the motion of the shock wave relative to the gas there correspond the estimates

$$|\partial(ur^2)/\partial r| \delta r < 0.2, \quad |\partial(v \sin \theta)/\partial \theta| \delta \theta < 0.3,$$

$$|\partial \rho / \partial r| \delta r < 3, \quad |\partial \rho / \partial \theta| \delta \theta < 0.1.$$

It is easy to see that the coefficients of the second derivatives

$$\begin{array}{ccccccc} \partial^2 \rho / \partial r^2, & \partial^2 \rho / \partial \theta^2, & \partial^2 u / \partial r^2, & \partial^2 u / \partial \theta^2, & & & \\ \partial^2 v / \partial r^2, & \partial^2 v / \partial \theta^2, & \partial^2 E / \partial r^2, & \partial^2 E / \partial \theta^2, & \partial^2 H / \partial \theta^2 & & \end{array}$$

will be positive, that is, this system of difference equations is stable.

4. Results of the calculations

As mentioned above we put the external boundary of the domain considered r_N equal to 1.1 a.u., and the internal boundary r_0 equal to 0.1 a.u. This makes it possible to consider dynamic processes in the interplanetary medium outside the regions of solar chromospheric flares, without going into the complex nature of their formation. As the scale for measuring distances we choose the distance from the sun to the earth's orbit, that is, $r_a = 1$ a.u. $= 1.495 \times 10^8$ km, as characteristic parameters of the flow we take the values u_0 , T_0 , ρ_0 and H_0 for $r = r_0$, corresponding to the data of observations: $u_0 = 339$ km/sec, $T_0 = 7.9 \times 10^5$ °K, $\rho_0 = 1.26 \times 10^3$ cm $^{-3}$, $H_0 = 4 \times 10^{-3}$ gauss.

Numerical calculations were performed for a flare characterized by the energy 0.6×10^{30} erg, velocity of the shock wave at the initial instant in the radial direction $u_B = 1500$ km/sec, semi-vertical angle of the cone in which the initial perturbation was concentrated, $\theta_* = 23^\circ$, initial thickness of the perturbed layer $\Delta r = 0.01$ a.u. The time interval in which the main energy of the flare was emitted was put equal to 0.1 hour. In performing the numerical calculations it was considered that the radial step $h = 0.005$ a.u., the angular step $\Delta\theta = 5^\circ 45'$, the angle θ being measured from a line connecting the centres of the sun and the earth.

To estimate the accuracy of the numerical calculations a trial calculation was performed with the radial step doubled. The relative variation of the results in the main parameters desired turned out not to exceed 14%, which testifies to an acceptable degree of accuracy.

Using Eq. (1.8) and Parker's model of the interplanetary magnetic field [1], we can calculate the parameters of the unperturbed magnetic field: $H_r = H_0 (r_0/r)^2$, $H_\theta = 0$, $H_\phi = H_0 r_0^2 \Omega \cos \theta_1 / u(r)r$. Here θ_1 is the angle between the polar axis and the plane of the ecliptic of the sun, $\Omega = 2.7 \times 10^{-6}$ sec $^{-1}$ is the angular velocity of the sun's rotation, and $u(r)$ is the velocity of the quiescent solar wind.

Considering that the lines of flow deviate little from the radial direction, using the freezing-in integral and the law of conservation of mass, it is easy to obtain an approximate expression for the azimuthal component of the intensity vector of the perturbed magnetic field:

$$H_\phi(r, \theta) = H_\phi \rho(r, \theta) / \rho_c(r, \theta),$$

where $\rho(r, \theta)$, $\rho_c(r, \theta)$ are respectively the densities of the perturbed and quiescent solar wind. The total strength of the interplanetary magnetic field H is given by $H = (H_r^2 + H_\phi^2)^{1/2}$.

Figure 2 shows the shape and leading front of the shock wave as it propagates through the interplanetary medium. The shock wave front for the case where the magneto-hydrodynamic effects are taken into account is shown by the continuous curves and reaches 1 a.u. for $\theta = 0^\circ$ after 60.4 hours. The dashed curves correspond to the shock wave front when the effect of the magnetic field is neglected. It is seen from Fig. 2 that the magnetic field delays the shock wave, the arrival time of its leading front at the earth's orbit for $\theta = 0^\circ$ being increased by 5 hours for the version

considered. It is easy to see that the shock wave does not penetrate into the exterior of the region bounded by a cone with semi-vertical angle $\theta = 46^\circ$.

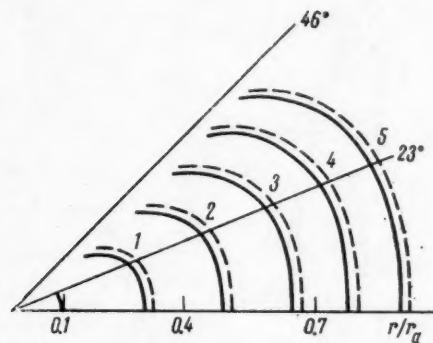


FIG. 2.

Position of the leading front of the shock wave: 1 is for $t = 10.5$ hours, 2 is for $t = 21$ hours, 3 is for $t = 31.5$ hours, 4 is for $t = 42$ hours, 5 is for $t = 52.5$ hours.

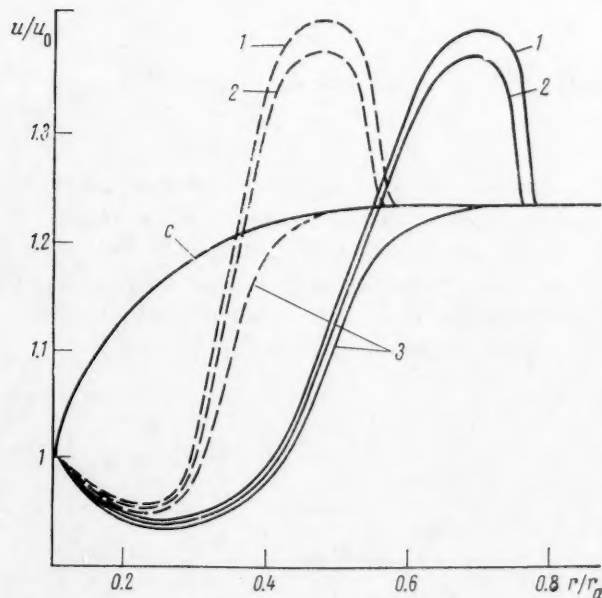


FIG. 3.

Velocities of the perturbed flow: 1 is for $\theta = 0^\circ$, 2 is for $\theta = 23^\circ$, 3 is for $\theta = 46^\circ$.

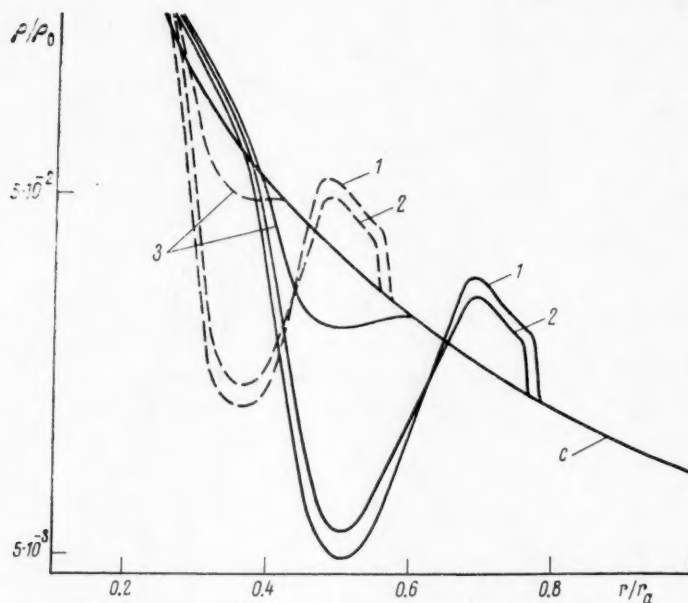


FIG. 4.

Densities of the perturbed flow: 1 is for $\theta = 0^\circ$, 2 is for $\theta = 23^\circ$,
3 is for $\theta = 46^\circ$.

The arrival time of the shock wave front at points of observation situated at the same distance from the sun, but at different angles to the axis of symmetry of the flare, will be different, as will also be the values of the velocities recorded (see Fig. 3) and the densities of the perturbed flow (see Fig. 4). The dashed curves here correspond to the distribution of the flow parameters at the instant $t_1 = 22$ hours, the continuous curves to the distribution of the velocity and density of the gas at the instant $t_2 = 36$ hours at the same angles.

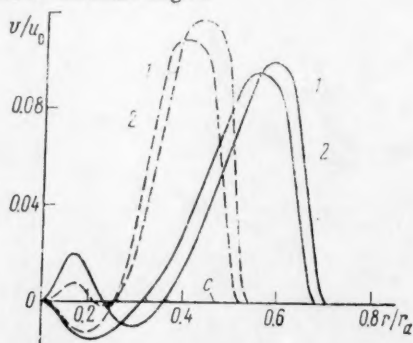


FIG. 5.

Variation of the tangential component v : 1 is for $\theta = 23^\circ$, 2 is for $\theta = 46^\circ$.

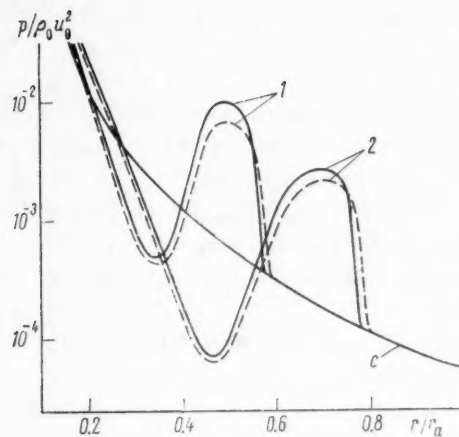


FIG. 6.

Distribution of pressure as a function of distance: 1 is for $t = 23.5$ hours, 2 is for $t = 37.4$ hours.

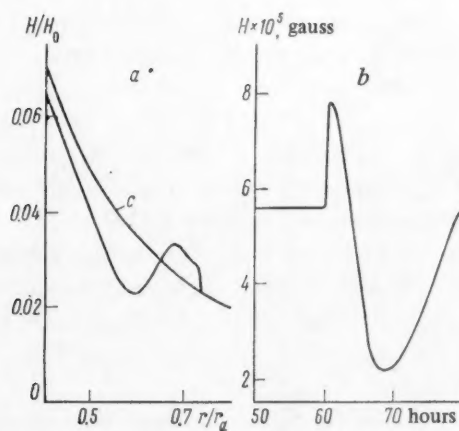


FIG. 7.

$$\theta = 0^\circ, \theta_1 = 15^\circ.$$

The curves with the symbol *c* in Figs. 3–7 show the distribution of the magnetohydrodynamic parameters in the stationary flow, that is, in the quiescent solar wind.

Figure 5 shows how the tangential component of the flow velocity v varies as a function of the distance r and the angle θ for two instants of time: $t_1 = 22$ hours (dashed curves) and $t_2 = 36$ hours (continuous curves). It is interesting to see that the velocity v attains values up to 70 km/sec at distances of 0.1 to 0.5 a.u. from the region occupied by the flare, but it makes no significant contribution to the total velocity of the flow at great distances from the sun (of the order

0.7 to 1 a.u.). It should also be mentioned that because of its three-dimensional nature the motion of the surface of the flow in the perturbed region has an extremely complex form, and in particular attention may be directed to the occurrence of zones with negative values of the tangential velocity component.

Figure 6 gives a comparison of the pressure distribution of the gas as a function of the distance r for $\theta = 0^\circ$ for two instants in the case where the magnetohydrodynamic effects are taken into account (continuous curves) and in the case where the effect of the magnetic field on the motion of the gas is neglected (dashed curves). Calculations did not reveal a significant effect of the magnetic field on the amplitude of the pressure, though the sum of the hydrodynamic and magnetic pressures may give an increase above the pressure ignoring the magnetic field of up to 10%.

The energy of the flare was calculated as indicated in [6].

Figure 7 shows the total strength of the magnetic field as a function of the distance r at the instant $t_1 = 36$ hours (Fig. 7, a) and of the time t measured from the beginning of the solar flare, if the observations are made from the point $r = r_a$ (Fig. 7, b). It is interesting to note, that the nature of the dependence of H on t is the same as that observed in reality by means of magnetometers mounted on cosmic equipment and satellites. On the passage of a shock wave the magnetometers register a sharp jump of the magnetic field strength, corresponding to the sudden beginning of a geomagnetic storm on the earth, and then a sharp fall of intensity in comparison with the magnetic field intensity of the quiescent solar wind, which corresponds to the so-called principal phase of a terrestrial geomagnetic storm.

It must be borne in mind that all the results relate to calculations for flares with energies of the order of 10^{31} ergs. The possibility is not excluded that for higher energies some of these results will require correction.

As a result of the numerical calculations performed the following parameter values of the quiescent solar wind at the earth's orbit were obtained: velocity of the wind $u = 417$ km/sec, density $\rho = 10 \text{ cm}^{-3}$, and strength of the total magnetic field $H = 5.54 \times 10^{-5}$ gauss. After the passage of the front of the shock wave through a point situated at a distance 1 a.u. from the sun for $\theta = 0^\circ$, which occurs 60.4 hours after the beginning of the development of the flare, the parameters of the perturbed solar wind acquire the following values: $u = 470$ km/sec, $\rho = 19 \text{ cm}^{-3}$, and $H = 7.75 \times 10^{-5}$ gauss.

The data of observations in interplanetary space, accumulated in recent decades, testify to the fact that at the earth's orbit the velocity and density of the quiescent solar wind are respectively of the order of 400 km/sec and 10 cm^{-3} , the magnetic field intensity varies within the limits 10^{-5} to 10^{-6} gauss, that the perturbation arrives at the earth's orbit 40 to 70 hours after the beginning of the solar flare, the velocity of the perturbed solar wind at the earth's orbit is of the order of 420 to 600 km/sec, and the density of the gas on the passage of the shock wave is increased by a factor of 2 to 3, as is also the magnetic field intensity. The results of the numerical solution discussed above agree with these data with an acceptable degree of accuracy, which permits us to regard the model of the propagation of the perturbation from a solar flare through the interplanetary gas, proposed in this paper, as completely satisfactory.

The numerical calculations were performed on the BESM-6 computer.

The author thanks V. P. Korobeinkova for his interest.

Translated by J. Berry.

REFERENCES

1. PARKER, E. *Dynamic processes in the interplanetary medium* (Dinamicheskie protsessy v mezhduplanetnoi srede), "Mir", Moscow, 1965.
2. KOROBENIKOV, V. P. and NIKOLAEV, Yu. M. Shock waves in the configuration of magnetic fields in interplanetary space. *Cosmic Electrodynamics*, 3, 1, 12-32, 1972.
3. SIMON, M. and AXFORD, W. J. Shock waves in interplanetary medium. *Planet. Space Sci.*, 14, 9, 901-908, 1966.
4. HUNDHAUSEN, A. J. and GENTRY, R. A. Numerical simulation of flare-generated disturbances in solar wind. *J. Geophys. Res.*, 74, 11, 2908-2919, 1969.
5. KOROBENIKOV, V. P. and SHIDLOVSKAYA, L. V. Numerical solution of problems of a flare in a moving gas. In: *Numerical methods of the mechanics of a continuous medium* (Chisl. metody mekhan. sploshnoi sredy). Inf. byull. Vol. 6, No. 4, "Nauka", Novosibirsk, 56-68, 1975.
6. SHIDLOVSKAYA, L. V. On the propagation of perturbations in interplanetary plasma caused by solar flares. *Dokl. Akad. Nauk SSSR*, 225, 2, 39-43, 1975.
7. BRANDT, J. and HODGE, P. *Solar System Astrophysics* (Astrofizika solnechnoi sistemy), "Mir", Moscow, 1967.
8. De YOUNG, D. S. and HUNDHAUSEN, A. J. Two-dimensional simulation of flare-associated disturbances in the solar wind. *J. Geophys. Res.*, 76, 10, 2245-2253, 1971.
9. BELOTSEKOVSKII, O. M. and DAVYDOV, Yu. M. A non-stationary "coarse particle" method for gas-dynamical computations. *Zh. vychisl. Mat. mat. Fiz.*, 11, 1, 176-207, 1971.
10. KULIKOVSKII, A. G. and LYUBIMOV, G. A. *Magnetohydrodynamics* (Magnitnaya gidrodinamika), Fizmatgiz, Moscow, 1962.

CONVERGENCE OF THE ITERATIVE PROCESS FOR THE QUASILINEAR HEAT TRANSFER EQUATION*

V. I. MASLYANKIN

Moscow

(Received 10 May 1976)

THE possibility of a numerical solution of the quasilinear heat-transfer equation by means of through-calculation difference schemes is discussed. It is shown that the convergence of the iterative process depends on the choice of the interpolation formula for the heat transfer coefficient. Approximate estimates of the regions of applicability of these formulas and the results of numerical calculations are presented.

1. Introduction

This paper is devoted to the numerical solution of the heat-conduction equation

$$\frac{\partial \varphi(u)}{\partial t} = \frac{\partial}{\partial x} \left[K(u) \frac{\partial u}{\partial x} \right] \quad (1)$$

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 209-216, 1977.

by means of homogeneous through-calculation difference schemes. As shown in [1], such schemes can be used to calculate generalized solutions of the temperature-wave type. The purpose of the paper is to analyze some interpolation formulas for the thermal conductivity $K(u)$, occurring in the difference operator approximating the right side of (1).

In sections 3 and 4 it is shown that on the introduction of an iterative process for solving the difference system of equations corresponding to (1), constraints arise on the applicability of some interpolation formulas for the thermal conductivity. But if these constraints are not applied, then the numerical solution may differ qualitatively from the actual solution. The derivatives made in sections 3 and 4 are confirmed by the numerical calculations presented in section 5.

We consider a heat wave propagating against a zero background. Let

$$K(0) = \varphi(0) = 0, \quad K(u) > 0, \quad \varphi'(u) > 0 \quad \text{for } u > 0,$$

$$\lim_{u \rightarrow 0} \left[\frac{K(u)}{\varphi'(u)} \right] = 0.$$

In this case the wave front propagates with finite velocity. Denoting the position of the front at the instant t by $\xi(t)$ (Fig. 1), we can obtain the following expressions for the velocity of the front [1]:

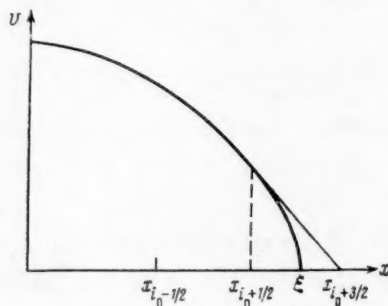


FIG. 1.

$$\frac{d\xi}{dt} = - \lim_{x \rightarrow \xi=0} \left[\frac{K(u)}{\varphi(u)} \frac{\partial u}{\partial x} \right]. \quad (2)$$

As examples in the calculations the analytic solution of Eq. (1) was used, which represents a running wave against a zero background (the thermal conductivity is taken in the form $K(u) = \kappa_0 u^\sigma$, and we assume $\varphi(u) \equiv u$):

$$u(t, x) = \begin{cases} [\sigma c \kappa_0^{-1} (ct + x_1 - x)]^{1/\sigma}, & x \leq x_1 + ct, \\ 0, & x_1 + ct \leq x, \end{cases} \quad (3)$$

where c is the velocity of motion of all the points of the profile, $c = \text{const}$. In the calculations we put $x_1 = 0$.

2. Scheme of the calculation. Interpolation formulas for the thermal conductivity

We suppose that $K(u) = \kappa_0 u^\sigma$, in which case (1) assumes the form

$$\frac{\partial \varphi(u)}{\partial t} = \frac{\partial}{\partial x} \left[\kappa_0 u^\sigma \frac{\partial u}{\partial x} \right]. \quad (4)$$

The boundary conditions for (4) have the form

$$u(t, 0) = \mu(t), \quad u(t, l) = \bar{\mu}(t).$$

Equation (4) is replaced by an implicit homogeneous difference scheme which is a modification of the scheme presented in [1]:

$$\varphi(v_i) - \varphi(\check{v}_i) = A_{i+1}(v_{i+1} - v_i) - A(v_i - v_{i-1}), \quad (5)$$

where $A_i = (\tau/h) \bar{K}_i$, and for the \bar{K}_i we use one of the three interpolation formulas described in [2] (the general method enabling these formulas to be obtained has been described previously in [3-6]):

$$\bar{K}_i = \frac{1}{h} K\left(\frac{v_{i-1} + v_i}{2}\right) \theta_i, \quad (6)$$

$$\bar{K}_i = \frac{2K(v_{i-1})K(v_i)}{h[K(v_{i-1}) + K(v_i)]} \theta_i, \quad (7)$$

$$\bar{K}_i = \frac{1}{2h} [K(v_{i-1}) + K(v_i)] \theta_i, \quad (8)$$

where $\theta_i = 1$ for $i = 1, 2, \dots, N-1$, $\theta_i = 0.5$ for $i = 0, N$. Quantities without a "halo" are computed at the step $j+1$, and quantities with a "halo" at the step j . The mesh is assumed to be uniform: $x_i = ih$, $0 \leq i \leq N$, $t^j = j\tau$.

The distinction from [1] consists of the fact that v_i refers to the half-integral point $i + 1/2$. We assume that $v_{-1/2} = v_0$, $v_{N+1/2} = v_N$. For this scheme stability occurs for any step τ .

The system of equations (5) for $i = 0, 1, \dots, N-1$ at each step $j+1$ is solved as in [1] by an iterative method. The resulting system of linear equations is solved by pivotal condensation (see, for example, [2]):

$$v_i^{[s+1]} = \alpha_{i+1} v_{i+1}^{[s+1]} + \beta_{i+1}, \quad (9)$$

where α_{i+1} and β_{i+1} are pivotal coefficients and s is the number of the iteration.

As the zeroth iteration values from the preceding step $v_i^{[0]} = \check{v}_i$ are chosen.

The condition for ending the iteration has the form

$$\max_{0 \leq i \leq N-1} |v_i^{[s+1]} - v_i^{[s]}| < \varepsilon + \varepsilon_1 v_i^{[s+1]}. \quad (10)$$

In all the calculations we assumed $\varepsilon_1 = 0$. For each example the actual number of iterations ν^j and the so-called "Courant relation"

$$\chi = \max [K(u) \tau / h^2], \quad (11)$$

characterizing the size of the time step were considered.

3. The difference running wave

As in [1] we will describe as difference running waves for Eq. (5) all the solutions of the equation

$$\sum_{k=1}^{\alpha} [A_{i+1}(v_{i+1} - v_i)]^{j-k+1} + \sum_{k=1}^{\beta} \varphi(v_{i+k}) = c_1, \quad (12)$$

where $\beta \geq 1$, $\alpha \geq 1$ are integers. The solutions of Eq. (12) preserve their profile, shifting after α steps (in j) through β computing intervals to the right, that is, $v_i^j = v_{i-\beta}^{j-\alpha}$. Therefore, the velocity of motion of all the points of the profile is constant and equals $c = \beta h / (\alpha \tau)$.

The solution of Eq. (12), being a wave moving relative to a quasi-zero background (see Fig. 1), can be selected as follows. Let $\xi_0 = x_{i_0 + 1/2}$ be the position of the difference front (we recall that the value v_{i_0} refers to the point $i_0 + 1/2$). We will consider that $v_{i_0+1} = \dots = v_{i_0+\beta} = \eta \geq 0$.

Putting $v_{i_0} \gg \eta$, the value of v_{i_0} can be determined from Eq. (2), assuming that the equation

$$-\frac{K(u)}{u} \frac{\partial u}{\partial x} = \frac{\beta h}{\alpha \tau}$$

holds at the point x_{i_0+1} (here and below we put $\varphi(u) \equiv u$):

$$-\frac{(h/\tau) A_{i_0+1}(v_{i_0+1} - v_{i_0})}{0.5(v_{i_0+1} + v_{i_0})} = \frac{\beta h}{\alpha \tau}.$$

From this we obtain

$$A_{i_0+1} \left\{ 1 - 2 \frac{\eta}{v_{i_0}} + O \left[\left(\frac{\eta}{v_{i_0}} \right)^2 \right] \right\} = 0.5 \beta / \alpha.$$

We will find all the values of v_i for $i < i_0$ from the equation

$$\sum_{h=1}^{\alpha} [A_{i+1} (v_{i+1} - v_i)]^{j-h+1} + \sum_{h=1}^{\beta} v_{i+h} = c_1,$$

where $c_1 = 0.5 \beta (\eta - v_{i_0})$.

We now consider the various forms of the interpolation formulas for the coefficient \bar{K}_i , $A_i = (\tau/h) \bar{K}_i$, putting $K(u) = \kappa_0 u^\sigma$, $\varphi(u) \equiv u$.

For (6), assuming $\sigma \eta / v_{i_0} \ll 1$, we have

$$v_{i_0} = 2 (2\sigma)^{-1/\sigma} \tilde{v}_{i_0} \left[1 + O \left(\frac{\eta}{v_{i_0}} \right) \right],$$

where \tilde{v}_{i_0} is the corresponding value on the analytic running wave (3).

For (8) we have

$$v_{i_0}^\sigma \left[1 + \left(\frac{\eta}{v_{i_0}} \right)^\sigma \right] \left(1 - 2 \frac{\eta}{v_{i_0}} \right) = \frac{ch}{\kappa_0},$$

or, considering that

$$v_{i_0} = \sigma^{-1/\sigma} \tilde{v}_{i_0} [1 + O(\eta/v_{i_0})],$$

For formula (7) for $\sigma \geq 1$ we have

$$v_{i_0} = \frac{2\eta}{1 - \tilde{v}_{i_0}^\sigma / 4\sigma \eta^\sigma}.$$

Since $v_{i_0} \gg \eta \geq 0$, then $\eta > (4\sigma)^{-1/\sigma} \tilde{v}_{i_0}$.

It is obvious from this that the "background" value of the temperature is at least comparable with the value at the front of the difference running wave. It is impossible to construct a difference running wave on the zeroth background for the interpolation formula (7).

We also note that since for large σ the expression $\sigma^{1/\sigma} \rightarrow 1$, then for the difference running wave defined by (6), the value at the front will be twice the actual value for large values of σ .

4. Convergence of the iterative process

We consider a temperature wave running along a quasi-zero background. Let $\varphi(u) \equiv u$, $K(u) = \kappa_0 u^\sigma$, $\sigma \geq 1$.

Let x_{i_0+1} be the position of the difference front. We will assume that $v_{i_0+1} = \dots = v_N$.

where $v_{i_0} \gg v_{i_0+1}$. In the subsequent analysis it is assumed that v_{i_0+1} is any sufficiently small background value. The analysis remains true for the case $v_{i_0+1} = 0$.

We consider a coefficient of the form (7), then

$$\bar{K}_{i_0+1} = \frac{2K(v_{i_0})K(v_{i_0+1})}{h[K(v_{i_0}) + K(v_{i_0+1})]}.$$

Since $v_{i_0} \gg v_{i_0+1}$ and $\sigma \geq 1$, then

$$\bar{K}_{i_0+1} = \frac{2}{h} K(v_{i_0+1}) - o\left[K(v_{i_0+1}) \frac{2}{h}\right].$$

Neglecting terms of the second order of smallness, from the pivotal formulas we obtain

$$\beta_{i_0+2} = \check{v}_{i_0+1} + \beta_{i_0+1} 2 \frac{\tau}{h^2} K_{i_0+1},$$

$$\Delta v_{i_0+1} = v_{i_0+1} - \check{v}_{i_0+1} = \beta_{i_0+1} 2 \frac{\tau}{h^2} K_{i_0+1}.$$

It is obvious (see (9)) that $\beta_i \leq v_0 = v_{\max}$. We put $\beta_{i_0+1} = v_{\max}$. Since the calculation is continued until condition (10) is satisfied, in order that the scheme be "sensitive", that the value of v_{i_0+1} be varied, it is necessary that Δv_{i_0+1} be greater than ϵ (that is, so that condition (10) will not be satisfied). Hence

$$v_{\max} \frac{2\tau}{h^2} K_{i_0+1} > \epsilon.$$

Using expression (11), we obtain

$$2\chi(v_{i_0+1}/v_{\max})^\sigma > \epsilon/v_{\max},$$

or

$$v_{i_0+1} > \left(\frac{\epsilon}{2\chi}\right)^{1/\sigma} v_{\max}^{1-1/\sigma}. \quad (13)$$

From this it is obvious that the background cannot be a zero one (for \bar{K}_i defined by formula (7)), moreover, the background value of the temperature cannot be less than some value, not always sufficiently small. It is interesting to note that on making the net finer, that is, decreasing the Courant ratio χ , we will have to increase the background value of the temperature.

We now consider a coefficient of the form (6). For it the background value of the temperature is unimportant, since in this case

$$\bar{K}_{i_0+1} \approx \frac{1}{h} K\left(\frac{v_{i_0}}{2}\right).$$

However, for a sufficiently great degree the σ in the thermal conductivity \bar{K}_{i_0+1} may be arbitrarily small. In this case, as before, we can write down the condition for which the scheme will work correctly:

$$v_{\max} \frac{\tau}{h^2} K_{i_0} 2^{-\sigma} > \epsilon,$$

or, using (11), we obtain

$$\sigma < \frac{\log_2(v_{\max}\chi/\varepsilon)}{1 - \log_2(v_{i_0}/v_{\max})}. \quad (14)$$

As in the case described above (that is, for formula (7)), the smaller χ , that is, the finer the mesh, the smaller the range of applicability of formula (6).

However condition (14) is too strong. If it is not satisfied, then thereby a value $v_{i_0+1} = v_N$ is defined at the point $x_{i_0+1/2}$. In this case the heating process is no longer described by Eqs. (3), the temperature v_{i_0} at the point $x_{i_0+1/2}$ increases, and, substituting the limiting value from (14) we obtain

$$\sigma < \log_2 \frac{\chi v_{\max}}{\varepsilon}. \quad (15)$$

If condition (15) is satisfied then as the value of v_{i_0} increases an instant may arrive when the condition of convergence of the iterations (10) for the points $x_{i_0+1/2}$ is violated. As a result a true numerical solution may be obtained. The corresponding numerical calculation is described below.

5. Examples of the numerical calculations

1. The following parameters of the calculation were chosen: $\varepsilon = 0.001$, $\sigma = 2$, $\kappa_0 = 0.5$, $c = 5$. Then from (3) the boundary conditions have the form $u(t, 0) = 10t^{1/2}$, $u(t, x_N) = 0$. The initial conditions for $t_0 = 0.1$ were chosen by solving (3). The calculation was performed up to $t_k = 0.2$ with step $\tau = 2 \cdot 10^{-1}$. The mesh was chosen sufficiently fine: $h = 0.02$, $N = 50$. In this case the Courant ratio (11) will be $\chi = 5.0$. From the inequality (13) we obtain that the background value of the temperature must be greater than $v^* \approx 0.024$.

In the calculations it was chosen equal to 10^{-5} . The analytic solution (continuous curves) and the results of the calculation (crosses) by formula (6) are shown in Fig. 2. Everywhere, apart from several nodes close to the front, the deviation of the calculated from the exact solution does not exceed 0.002, the number of iterations $\nu \leq 3$. Calculation by formula (8) gives practically the same results.

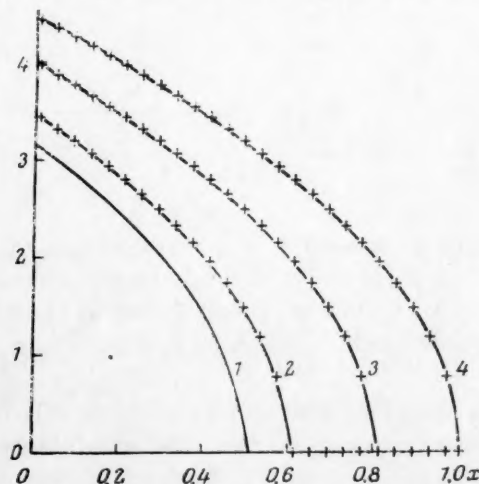


FIG. 2.

Solution: 1 is for $t_0 = 0.10$, 2 is for $t_1 = 0.12$, 3 is for $t_2 = 0.16$, 4 is for $t_k = 0.20$;

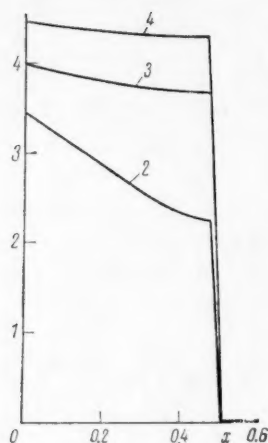


FIG. 3.

Solution: 1 is for $t_0 = 0.10$, 2 is for $t_1 = 0.12$, 3 is for $t_2 = 0.16$, 4 is for $t_k = 0.20$;

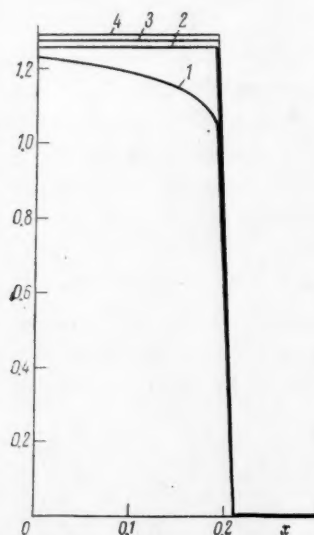


FIG. 4.

The results of the same calculation by formula (7) are shown in Fig. 3. Since the background value of the temperature is much less than ν^* , then by the previous analysis the convergence condition (10) is satisfied at the node closest to the front. Since the number of iterations $\nu \leq 3$, the value of the temperature at this node is practically unchanged and the front remains immobile.

The spatial step was changed in the calculations, the steps $h = 0.04$ and 0.01 were chosen. The results of these calculations are practically the same as the results obtained for $h = 0.02$.

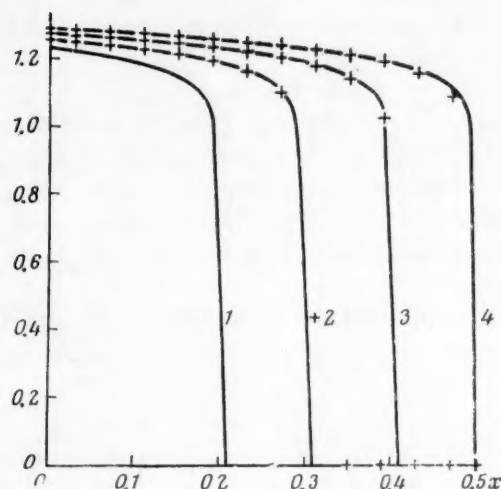


FIG. 5.

Solution: 1 is for $t_0 = 0.10$, 2 is for $t_1 = 0.15$, 3 is for $t_2 = 0.20$, 4 is for $t_k = 0.25$.

In the case $h = 0.02$ a calculation was performed in which the initial instant t_0 was so chosen that condition (14) was violated. It is easy to see that condition (15) is satisfied. The calculation was carried out using formula (6), and the numerical solution obtained is identical with that considered above (where $t_0 = 0.1$ was chosen). This confirmed the inferences made above concerning formulas (14) and (15).

2. For a calculation illustrating the case where formula (6) does not work the following parameters were chosen: $\varepsilon = 0.001$, $\sigma = 20$, $\kappa_0 = 1/8$, $c = 2$. The boundary conditions from (3) are of the form $u(t, 0) = (640 \cdot t)^{0.05}$, $u(t, x_N) = 0$.

The Courant ratio is $\chi = 16$, $h = 0.02$, $N = 25$, $\tau = 2 \cdot 10^{-4}$. The initial conditions for $t_0 = 0.10$ were chosen from the solution of (3). The calculation was performed up to $t_k = 0.25$.

By the inequality (15), we have the condition $\sigma < \sigma^* \approx 14$, which is obviously not satisfied.

The results of the calculation by formula (6) are given in Fig. 4. The number of iterations $\nu \leq 3$, therefore the value of the temperature at the front cannot be varied in practice and the front is immobile.

The analytic solution (continuous curves) and the results of calculation (crosses) by formula (8) are shown in Fig. 5. From $t_0 = 0.10$ to $t_1 = 0.15$ the number of iterations $\nu \leq 11$, from $t_1 = 0.15$ to $t_k = 0.25$ the number of iterations $\nu \leq 3$. Everywhere, apart from some nodes close to the front, the deviation of the calculated from the exact solution does not exceed 0.003.

6. Conclusion

From the analysis of formulas (6)–(8) it is obvious that formula (7) is practically useless for calculating temperature waves propagating along a fairly small background, and for calculating processes with a large temperature drop.

Formula (6) is inappropriate for large degrees ($\chi \leq 20$) in the thermal conductivity. However, even in the case of real values of the exponent σ (for example, $\sigma = 2$) for sufficiently fine meshes ($\sigma \geq 20$) calculation by formula (8) is advantageous, since it requires less computer time. For example, the calculations carried out in section 5, subsection 1 for $h = 0.02$ required 30% more time when formula (6) was used than when (8) was used.

For coarse meshes formula (6) is not suitable, it enables a more accurate solution to be obtained, and as shown by calculations, on coarse meshes it requires less computer time than formula (8).

I express my thanks to A. A. Samarskii for considerable assistance and for his continued interest.

Translated by J. Berry.

REFERENCES

1. SAMARSKII, A. A. and SOBOL, I. M. Examples of the numerical calculation of temperature waves, *Zh. vychisl. Mat. mat. Fiz.*, 3, 4, 702–719, 1963.
2. SAMARSKII, A. A. *Introduction to the theory of difference schemes* (Vvedenie v teoriyu raznostnykh skhem), "Nauka", Moscow, 1971.
3. SAMARSKII, A. A. Equations of parabolic type with discontinuous coefficients and difference methods for their solution. *Proceedings of the All-Union Conference on Differential Equations* (Tr. Vses. sovmeshchaniya po differents. ur-niyam) (Erevan, November 1958), 148–160, Izd-vo Akad. Nauk Arm SSR, Erevan, 1960.
4. SAMARSKII, A. A. On the convergence and accuracy of homogeneous difference schemes for one-dimensional and multidimensional parabolic equations. *Zh. vychisl. Mat. mat. Fiz.*, 2, 4, 603–634, 1962.
5. SAMARSKII, A. A. Homogeneous difference schemes on non-uniform nets for equations of the parabolic type. *Zh. vychisl. Mat. mat. Fiz.*, 3, 2, 266–298, 1963.
6. SAMARSKII, A. A. and FRYAZINOV, I. V. On the convergence of difference schemes for the heat-conduction equation with discontinuous coefficients. *Zh. vychisl. Mat. mat. Fiz.*, 1, 5, 806–824, 1961.

LOCAL ALGORITHMS ON YABLONSKII SCHEMES*

A. V. KABULOV

Moscow

(Received 17 May 1974)

LOCAL algorithms on information processing systems are discussed. It is shown that the concept of best local algorithm for the introduction of a system of neighborhoods extends to a class of control systems.

**Zh. vychisl. Mat. mat. Fiz.*, 17, 1, 217–225, 1977.

In this paper we consider local algorithms for calculating information [1] on the elements of control systems [2].

It will be shown that for the elements of such systems a system of neighborhoods can be introduced in such a way that the neighborhoods satisfy fundamental axioms.

It will also be shown that for neighborhoods in control systems and predicates characterizing the properties of elements in systems, the concept of best local algorithm, theorems of the existence of the best algorithm and the property of partial ordering in the class of best algorithms are preserved. All these results are obtained in section 2 of the present paper.

In section 1 the fundamental definitions of the theory of ordering systems, such as networks, memory, elements, subschemes and schemes are introduced. The interaction between the elements of ordering systems is described, and brief characteristics of their functioning are given. Section 1 is based on [2].

In the exposition of the results on the theory of local algorithms it is assumed that the reader is familiar with the fundamental definitions and theorems of this theory in [1].

1. Control systems

A strict definition of the class of information processing systems, or as the author calls them, control systems, was given in [2]. We consider below an essential element of control systems — schemes.

Following [2] we introduce successively the definitions of networks, memory, elements and subscheme.

The synthesis of these concepts gives the principle object of our investigation—schemes.

1. *Networks.* Let $\mathfrak{M} = \{a_v\}$ be a set of distinct objects a_v , possessing power m . Also let $E_0, E_i, i \geq 1$, be groups of objects (by groups of objects we mean an unordered collection of objects, in which the repetition of them is possible) of the set \mathfrak{M} , possessing a power taking into account repetitions, e_0 and e_i , respectively. We assume that the subscript runs through a set of natural numbers of power h , where the different subscripts may correspond to identical groups.

Definition 1. The set \mathfrak{M} with the assigned aggregate of groups E_0, E_1, \dots is called a network and is denoted by $\mathfrak{M}(E_0, E_1, \dots)$, if $|E_0| \equiv \bigcup_{1 \leq i \leq h} |E_i|$. Here the symbol $|E|$ denotes

the set of all the objects of the group E . The objects occurring in the composition of the set \mathfrak{M} , are called vertices of the network, and the objects of the group E_0 are called poles of the network.

In the study of real control systems a fundamental role is played by networks for which the numbers $m, h, e_i, i = 0, 1, \dots, h$, are natural numbers. Such networks are called finite.

Let $\mathfrak{M}(E_0, E_1, \dots, E_h)$ be a finite network and $E_i = (a_1^i, \dots, a_{l_i}^i)$, where $a_j^i \in \mathfrak{M}$, $i = 0, 1, \dots, h$. With each group E_i we associate a circle in three-dimensional space, and with the objects $a_1^i, \dots, a_{l_i}^i$ of the group E_i we associate rays emerging from this circle.

With the group E_0 we associate in three-dimensional space points, each possessing one ray, each of which corresponds to one of the objects a_1^0, \dots, a_l^0 . We assume that all the rays corresponding to the same objects of the set \mathfrak{M} , are connected to each other. The diagram obtained as a result of the construction (Fig. 1), is called the geometrical realization of the network, if: a) each pair of circles occurring in the diagram has no common points; b) the junctions of the rays corresponding to different vertices a_i and a_j have no common points.

2. Memory, elements, elementary subschemes.

Definition 2. The set $\chi = \{X_\alpha\}$ of distinct objects is called a memory; the objects X_α are called cells.

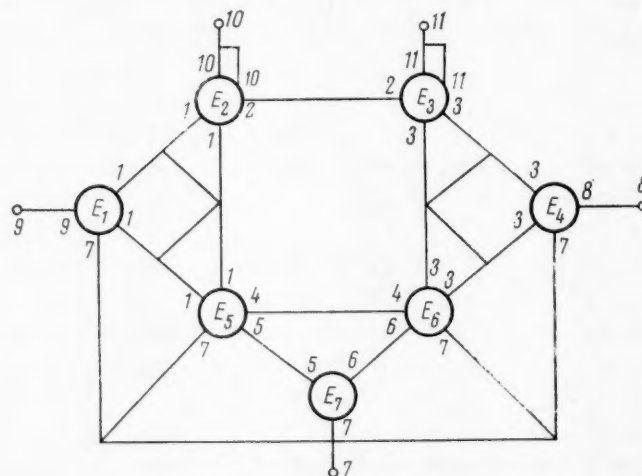


FIG. 2.

Semantically a memory is understood as a receptacle for the storage and memorization of information.

Definition 3. The symbol $C_\alpha(, ,)$, possessing three empty places and a certain number of poles, is called an element, if there is indicated: the number c_α of poles which the symbol C_α has, and cardinal numbers $u_\alpha, v_\alpha, w_\alpha$, corresponding to the first, second and third empty places.

Here we must note that the numbers $u_\alpha, v_\alpha, w_\alpha$ may be zeros. Below the elements will be linked, on the one hand with networks, on the other hand with the memory. Indeed, the poles of the element C_α will be put in a one-to-one relation with the objects of the group E_i . The latter is possible only on condition that $c_\alpha = e_i$. As for the empty places of the symbol $C_\alpha(, ,)$, groups $X_\alpha, Y_\alpha, Z_\alpha$ of cells of some memory χ will be substituted in them; and then the powers of the groups $X_\alpha, Y_\alpha, Z_\alpha$ will equal $u_\alpha, v_\alpha, w_\alpha$ respectively.

Let χ be some memory, E_α a group of objects of the set \mathfrak{M} and $C_\alpha(, ,)$ an arbitrary element possessing c_α poles, with whose empty places are associated the numbers $u_\alpha, v_\alpha, w_\alpha$.

Definition 4. The symbol $C_\alpha = C_\alpha(X^\alpha, Y^\alpha, Z^\alpha)$ is called an elementary subscheme on the

memory χ , if with the poles of the element $C_\alpha(, ,)$ are associated respectively objects of the group E_α possessing the power c_α , and for the groups $X^\alpha, Y^\alpha, Z^\alpha$ of the memory χ , possessing respectively the power $u_\alpha, v_\alpha, w_\alpha$, the condition

$$(|X^\alpha| \cup |Y^\alpha|) \cap |Z^\alpha| = \emptyset.$$

is satisfied.

Here the group Z^α defines those cells of the memory χ , which are rigidly connected with the given elementary subscheme, these cells containing both information necessary for the operation of the elementary subscheme, and also information arising as a result of its operation. The group X^α selects those cells of the memory $\chi \setminus Z^\alpha$, containing information necessary for the operation of the elementary subscheme C_α . Finally, the group Y^α fixes those cells of the memory $\chi \setminus Z^\alpha$, in which information arrives which has emerged thanks to the operation of the elementary subscheme.

Definition 5. Let C_α be an elementary subscheme on the memory χ . We will call the set Z^α and $\chi \setminus Z^\alpha$, respectively, the internal and external memory of the elementary subscheme C_α .

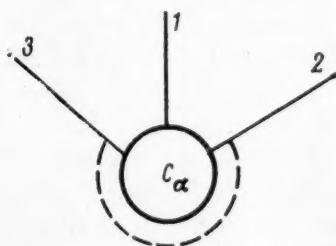


FIG. 2.

1 corresponds to a_{i1}^α , 2 corresponds to a_{i2}^α , 3 corresponds to a_{is}^α

For clarity (Fig. 2) we will provisionally represent elementary subschemes as a circle from which emerge c_α numbered rays (objects of the group E_α), with the symbol C_α at the centre.

3. Schemes.

Definition 6. Let σ_0 be a set of elementary subschemes on the memory χ and $\mathfrak{M}(E_0, E_1, \dots)$ be some network. The symbol $\mathfrak{M}(E_0, C_{\alpha_1}, C_{\alpha_2}, \dots)$ is called a scheme, if it is obtained as the result of substituting into the network $\mathfrak{M}(E_0, E_1, \dots)$ in a place of the groups E_1, E_2, \dots the elementary subschemes $C_{\alpha_i} = C_{\alpha_i}^{E_i}(X^{\alpha_i}, Y^{\alpha_i}, Z^{\alpha_i})$, $C_{\alpha_2} = C_{\alpha_2}^{E_2}(X^{\alpha_2}, Y^{\alpha_2}, Z^{\alpha_2}), \dots$, where $e_i = c_{\alpha_i}$, $i = 1, 2, \dots$, and the poles of the elementary subscheme $C_{\alpha_i} = C_{\alpha_i}^{E_i}(X^{\alpha_i}, Y^{\alpha_i}, Z^{\alpha_i})$ are put in a definite correspondence with the vertices of the group E_i , $i = 1, 2, \dots$

In particular, $\mathfrak{M}(E_0, C_\alpha) = C_\alpha$, where $E_0 = E_1 = E_\alpha$.

Definition 7. The sets $Z = \bigcup |Z^{\alpha_i}|$ and $\chi \setminus Z$ are called the internal and external memory of the scheme $\mathfrak{M}(E_0, C_{\alpha_1}, C_{\alpha_2}, \dots)$, respectively.

Schemes constructed on networks are conveniently represented geometrically. For this purpose it is sufficient in the geometrical image of the network to write the representation of the elementary subscheme C_{α_i} in the circle representing the group E_i .

2. Majorant local algorithms on Yablonskii schemes

1. Neighborhoods of the elements of control systems. Let a network $\mathfrak{M}(E_0, E_1, \dots)$ be given with a set of poles E_0 and a set of groups E_1, E_2, \dots .

Definition 8. The principal neighborhood of the first order $S_1(E_i, \mathfrak{M})$ of the group E_i of the network $\mathfrak{M}(E_0, E_1, \dots)$ is defined as all the groups $E_\alpha \in \mathfrak{M} | E_i \cap E_\alpha \neq \emptyset$.

Let the principal neighborhood of the $(k-1)$ -th order $S_{k-1}(E_i, \mathfrak{M})$ of the group E_i of the network $\mathfrak{M}(E_0, E_1, \dots)$ have been defined.

Definition 9. The principal neighborhood of the k -th order $S_k(E_i, \mathfrak{M})$ of the group E_i of the network $\mathfrak{M}(E_0, E_1, \dots)$ is defined as all the groups E_α of \mathfrak{M} , for which one of the following conditions is satisfied: 1) $E_\alpha \cap E_\beta \neq \emptyset$, $E_\beta \in S_{k-1}(E_i, \mathfrak{M})$ 2) $E_\alpha \subseteq \bigcup_k E_k$, where E_k satisfies condition 1).

If a finite network is given, then we know that for it there exists its own geometrical realization. Therefore the principal neighborhood of the k -th order for a group of a network can be defined as the principal neighborhood of the k -th order for the vertices of the graph corresponding to the given network.

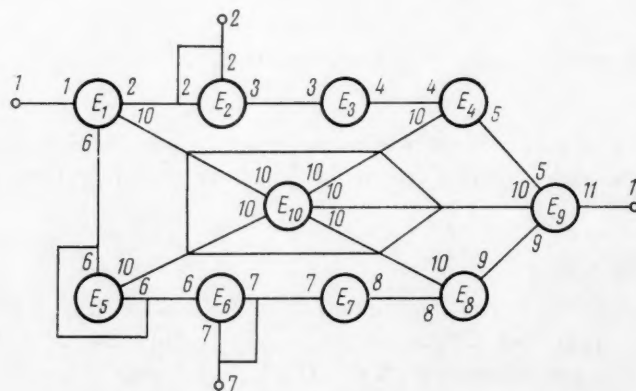


FIG. 3.

Example. Let $\mathfrak{M} = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11)$ be the network $\mathfrak{M}(E_0, E_1, \dots, E_{10})$, see Fig. 3. For it

$$\begin{array}{lll}
E_0 = (1, 2, 7, 11) & E_4 = (4, 5, 10) & \\
E_1 = (1, 2, 6, 10) & E_5 = (6, 6, 10) & E_8 = (8, 9, 10) \\
E_2 = (2, 2, 3) & E_6 = (6, 7, 7) & E_9 = (5, 9, 10, 11) \\
E_3 = (3, 4) & E_7 = (7, 8) & E_{10} = (10, 10, 10, 10, 10)
\end{array}$$

The principal neighborhoods:

$$\begin{aligned}
S_1(E_1, \mathfrak{M}) &= (E_1, E_0, E_2, E_5, E_{10}), \\
S_1(E_{10}, \mathfrak{M}) &= (E_{10}, E_1, E_4, E_5, E_8, E_9), \\
S_2(E_1, \mathfrak{M}) &= S_2(E_{10}, \mathfrak{M}) = \mathfrak{M}.
\end{aligned}$$

Let $\mathfrak{M}(E_0, C_{\alpha_1}, C_{\alpha_2}, \dots)$ be a scheme on the memory χ , and let σ_0 be a set of subschemes of the scheme \mathfrak{M} .

Definition 10. The principal neighborhood of the first order $S_1(C_{\alpha_i}, \mathfrak{M})$ of the subscheme C_{α_i} of the scheme \mathfrak{M} is defined as the aggregate of the subschemes C_{α_k} of \mathfrak{M} such that $C_{\alpha_i} \cap C_{\alpha_k} \neq \emptyset$ and $(|Y^{\alpha_i}| \cap (|X^{\alpha_k}| \cup |Z^{\alpha_k}|) = \emptyset) \vee (|Y^{\alpha_k}| \cap (|X^{\alpha_i}| \cup |Z^{\alpha_i}|) \neq \emptyset)$.

Let the principal neighborhood of the $(k-1)$ -th order of the subscheme C_{α_i} of the scheme \mathfrak{M} have been defined.

Definition 11. The principal neighborhood of the k -th order $S_k(C_{\alpha_i}, \mathfrak{M})$ of the subscheme C_{α_i} in \mathfrak{M} is defined as the aggregate of all the subschemes C_{α_k} of \mathfrak{M} , for which one of the following conditions is satisfied:

$$\begin{aligned}
C_{\alpha_k} \cap C_{\alpha_j} \neq \emptyset, \text{ where } C_{\alpha_j} \in S_{k-1}(C_{\alpha_i}, \mathfrak{M}), \\
(|Y^{\alpha_j}| \cap (|X^{\alpha_k}| \cup |Z^{\alpha_k}|) \neq \emptyset) \vee (|Y^{\alpha_k}| \cap (|X^{\alpha_j}| \cup |Z^{\alpha_j}|) = \emptyset);
\end{aligned} \tag{1}$$

$$\begin{aligned}
C_{\alpha_k} \subseteq \bigcup_j C_{\alpha_j}, \\
((|X^{\alpha_k}| \cup |Z^{\alpha_k}|) \subseteq \bigcup_j |Y^{\alpha_j}|) \vee (|Y^{\alpha_k}| \subseteq \bigcup_j (|X^{\alpha_j}| \cup |Z^{\alpha_j}|)),
\end{aligned} \tag{2}$$

where C_{α_j} satisfies condition 1).

In this definition of the principal neighborhood of the k -th order of a subscheme proximity in the network (in the topology) and in the memory is considered simultaneously. But this is not always the case. We consider particular cases.

Proximity in the network.

Definition 12. The principal neighborhood of the first order $S_1(C_{\alpha_i}, \mathfrak{M})$ of the subscheme C_{α_i} of the scheme \mathfrak{M} is defined as the aggregate of subschemes C_{α_k} of \mathfrak{M} , for which the following condition is satisfied:

$$C_{\alpha_i} \cap C_{\alpha_k} \neq \emptyset.$$

Let the principal neighborhood of the $(k-1)$ -th order of the subscheme C_{α_i} of the scheme \mathfrak{M} have been defined.

Definition 13. The principal neighborhood of the k -th order $S_k(C_{\alpha_i}, \mathfrak{M})$ of the subscheme C_{α_i} of the scheme \mathfrak{M} is defined as the aggregate of all the subschemes C_{α_k} of \mathfrak{M} , for which one of the following conditions is satisfied:

- 1) $C_{\alpha_k} \cap C_{\alpha_j} \neq \emptyset$, where $C_{\alpha_j} \in S_{k-1}(C_{\alpha_i}, \mathfrak{M})$;
- 2) $C_{\alpha_k} \subseteq \bigcup_j C_{\alpha_j}$, where C_{α_j} satisfies condition 1).

Proximity in the memory.

Definition 14. The principal neighborhood of the first order $S_1(C_{\alpha_i}, \mathfrak{M})$ of the subscheme C_{α_i} of \mathfrak{M} is defined as the aggregate of all the subschemes C_{α_k} of \mathfrak{M} for which the following condition is satisfied:

$$(|Y^{\alpha_i}| \cap (|X^{\alpha_k}| \cup |Z^{\alpha_k}|)) \neq \emptyset \vee (|Y^{\alpha_k}| \cap (|X^{\alpha_i}| \cup |Z^{\alpha_i}|)) \neq \emptyset.$$

Let the principal neighborhood of the $(k-1)$ -th order of the subscheme C_{α_i} have been defined.

Definition 15. The principal neighborhood of the k -th order $S_k(C_{\alpha_i}, \mathfrak{M})$ of the subscheme C_{α_i} of \mathfrak{M} is defined as the aggregate of all the subschemes C_{α_k} of \mathfrak{M} , for which one of the following conditions is satisfied:

- 1) $(|Y^{\alpha_i}| \cap (|X^{\alpha_k}| \cup |Z^{\alpha_k}|)) \neq \emptyset \vee (|Y^{\alpha_k}| \cap (|X^{\alpha_i}| \cup |Z^{\alpha_i}|)) = \emptyset$,
where $C_{\alpha_j} \in S_{k-1}(C_{\alpha_i}, \mathfrak{M})$;
- 2) $((|X^{\alpha_k}| \cup |Z^{\alpha_k}|) \subseteq \bigcup_j |Y^{\alpha_j}|) \vee (|Y^{\alpha_k}| \subseteq (|X^{\alpha_i}| \cup |Z^{\alpha_i}|))$,

where C_{α_j} satisfies condition 1).

Below in the proofs of the theorems we will consider proximity simultaneously in the network and in the memory.

Let a set of subschemes $\sigma = \{C_\alpha\}$ of the scheme \mathfrak{M} , have been defined, and let $\{\sigma\}$ be a family of the set of subschemes. Let $S_i(C_\alpha, \mathfrak{M})$ be the principal neighborhood of the i -th order of the subscheme C_α in the set $\sigma_i \in \{\sigma\}$.

Theorem 1

The neighborhoods $S_1(C_\alpha, \sigma_i), S_2(C_\alpha, \sigma_i), \dots, S_k(C_\alpha, \sigma_i)$ are special neighborhoods.

The proof is by induction with respect to k .

1. Let $S_1(C_\alpha, \sigma_i) \subseteq \sigma_i \cap \sigma_\emptyset$, $S_1(C_\alpha, \sigma_\emptyset) \subseteq \sigma_i \cap \sigma_\emptyset$. The neighborhood $S_1(C_\alpha, \sigma_i)$ is composed of all the subschemes C_{α_i} such that $C_\alpha \cap C_{\alpha_i} \neq \emptyset$.

$$(|Y^\alpha| \cap (|X^{\alpha_i}| \cup |Z^{\alpha_i}|)) \neq \emptyset \vee (|Y^{\alpha_i}| \cap (|X^\alpha| \cup |Z^\alpha|)) \neq \emptyset.$$

All such C_{α_i} occur in $S_1(C_\alpha, \sigma_\emptyset)$, since $S_1(C_\alpha, \sigma_\emptyset) \subseteq \sigma_\emptyset$. Therefore $S_1(C_\alpha, \sigma_i) \subseteq S_1(C_\alpha, \sigma_\emptyset)$. The converse inclusion is proved similarly. Consequently,

$$S_1(C_\alpha, \sigma_f) = S_1(C_\alpha, \sigma_\Phi).$$

2. Let $S_{h-1}(C_\alpha, \sigma_f) = S_{h-1}(C_\alpha, \sigma_\Phi)$, $S_h(C_\alpha, \sigma_f) \subseteq \sigma_f \cap \sigma_\Phi$, $S_h(C_\alpha, \sigma_\Phi) \subseteq \sigma_f \cap \sigma_\Phi$. We show that $S_h(C_\alpha, \sigma_f) = S_h(C_\alpha, \sigma_\Phi)$. The set $S_h(C_\alpha, \sigma_f)$ is composed of subschemes C_{α_j} , such that:

1) either C_{α_j} has a non-empty intersection with C_{α_i} , where $C_{\alpha_i} \in S_{h-1}(C_\alpha, \sigma_f)$, and

$$(|Y^{\alpha_j}| \cap (|X^{\alpha_i}| \cup |Z^{\alpha_i}|)) = \emptyset \vee (|Y^{\alpha_i}| \cap (|X^{\alpha_j}| \cup |Z^{\alpha_j}|)) = \emptyset;$$

since $S_{h-1}(C_\alpha, \sigma_f) = S_{h-1}(C_\alpha, \sigma_\Phi)$, then $C_{\alpha_i} \in S_{h-1}(C_\alpha, \sigma_\Phi)$;

2) or C_{α_j} is contained in the sum of subschemes C_{α_i} such that C_{α_i} satisfies condition 1).

But all the C_{α_i} occur in $S_h(C_\alpha, \sigma_\Phi)$, and therefore C_{α_j} occurs in $S_h(C_\alpha, \sigma_\Phi)$ and consequently $S_h(C_\alpha, \sigma_f) \subseteq S_h(C_\alpha, \sigma_\Phi)$.

The converse inclusion is proved similarly. The theorem is proved.

2. The best local algorithm for elements of control systems. Let the memory χ and the set of elementary subschemes σ be fixed. We will consider the scheme

$$\Sigma_i = \mathfrak{M}_i(E_0, C_{T1}, C_{T2}, \dots).$$

Let the information vector $\bar{\alpha} = (\alpha_1, \dots, \alpha_l)$, be given, where $\alpha_i \in \{0, 1, \Delta\}$, $i = 1, 2, \dots, l$.

Let $\sigma = \{C_{\gamma_1}, \dots, C_{\gamma_l}\}$. We call the set $\sigma^* = \{C_{\gamma_1}^{\alpha_1}, \dots, C_{\gamma_l}^{\alpha_l}\}$, where

$$C_{\gamma_1}^{\alpha_1} = C_{\gamma_1}^{E_1^{\alpha_1}, \dots, \alpha_{l1}}(X^{\gamma_1}, Y^{\gamma_1}, Z^{\gamma_1}), \text{ a } C_{\gamma_l}^{\alpha_l} = C_{\gamma_l}^{E_l^{\alpha_l}, \dots, \alpha_{ll}}(X^{\gamma_l}, Y^{\gamma_l}, Z^{\gamma_l}),$$

permissible for σ .

We call the class $M^* = I(\sigma)$ of all the sets σ^* permissible for σ the information class of the set σ with respect to certain predicates P_1, \dots, P_l .

The neighborhood $S(C_\gamma, \sigma)$ determines the neighborhood $S(C_\gamma^!, \sigma^*)$, where

$$C_\gamma^! = C_\gamma^{E_\gamma^{\alpha_1}, \dots, \alpha_l}(X^\gamma, Y^\gamma, Z^\gamma), \quad \sigma^* \in I(\sigma).$$

Let the functions $\varphi_1, \dots, \varphi_l$ and the algorithm A , defined by the system of predicates P_1, \dots, P_l be given, and let the class of algorithms π_s with the same memory be given, where the domain of definition is the same for all the algorithms:

$$\pi_s = \{A_s, \varphi_1, \dots, \varphi_l, P_1, \dots, P_l, S\}.$$

Theorem 2

For every class π_s of locally equal algorithms with the same memory there exists a majorant algorithm.

Proof. Let $I(\sigma)$ be the information class of the set σ , where $\sigma \in \{\sigma\}$. We consider the set

$$M^* = \bigcup_{\sigma} I(\sigma).$$

Let $C_{\gamma}^1 \in \sigma^*$, where $\sigma^* \in M^*$, and $S(C_{\gamma}^1, \sigma)$ is the neighborhood of the subscheme C_{γ}^1 in σ^* .

We select in M^* all the sets σ_{α}^* such that $C_{\gamma}^1 \in \sigma_{\alpha}^*$ and $S(C_{\gamma}^1, \sigma^*) \supseteq S(C_{\gamma}^1, \sigma_{\alpha}^*)$. We denote the aggregate of all the sets σ_{α}^* by

$$M_s(C_{\gamma}^1). \quad (1)$$

The aggregate of sets σ such that in (1) there exists a σ^* from $I(\sigma)$ will be denoted by $M_s(C_{\gamma})$. We introduce the function φ_i^0 , $i=1, 2, \dots, l$, as follows:

$$\varphi_i^0(C_{\gamma}, \alpha_1, \dots, \alpha_l, S, \sigma^*) = \begin{cases} (\alpha_1, \dots, \alpha_l), & \text{if } \alpha_i \in \{0, 1\}; \\ (\alpha_1, \dots, \alpha_{i-1}, \gamma, \alpha_{i+1}, \dots, \alpha_l), & \text{if for} \\ \text{all } \sigma \in M_s(C_{\gamma}) \text{ there is satisfied the} \\ \text{relation } P_i(C_{\gamma}, \sigma) = \gamma, \gamma \in \{0, 1\}; \\ (\alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_l), & \\ \text{if in } M_s(C_{\gamma}) \exists \sigma_1 \text{ and} \\ \sigma_2 \mid P_i(C_{\gamma}, \sigma_1) \neq P_i(C_{\gamma}, \sigma_2). \end{cases}$$

We prove the monotonicity of the functions φ_i^0 , $i=1, 2, \dots, l$. Let $S_1 = S(C_{\gamma}^1, \sigma_1^*)$, $S_2 = S(C_{\gamma}^1, \sigma_2^*)$, then

$$M_{s_1}(C_{\gamma}) \supseteq M_{s_2}(C_{\gamma}). \quad (2)$$

Indeed, if $\sigma \in M_{s_2}(C_{\gamma})$, then a σ^* exists such that $\sigma^* \in M_{s_2}(C_{\gamma}^{E_{\gamma}^{\beta_1, \dots, \beta_l}}(X^{\gamma}, Y^{\gamma}, Z^{\gamma}))$.

We replace in σ^* some markers γ , $\gamma \in \{0, 1\}$, by Δ so as to obtain the equation $S(C_{\gamma}^1, \sigma_1^*) \supseteq S(C_{\gamma}^1, \bar{\sigma}^*)$; it is obvious that $\sigma^* \in I(\sigma)$ and $\bar{\sigma}^* \in M_{s_1}(C_{\gamma}^1)$, but then $\sigma \in M_{s_1}(C_{\gamma})$. The inclusion (2) is proved.

1. Let $\varphi_i^0(C_{\gamma}, \alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_l, S, \sigma_1^*) = (\alpha_1, \dots, \alpha_{i-1}, \gamma, \alpha_{i+1}, \dots, \alpha_l)$, $\gamma \in \{0, 1\}$. It follows from the definition of φ_i^0 that the equation $P_i(C_{\gamma}, \sigma) = 1$ is satisfied for all σ from $M_{s_1}(C_{\gamma})$. But then from (2) we obtain that the equation $P_i(C_{\gamma}, \sigma') = \gamma$ is also satisfied for all σ' from $M_{s_2}(C_{\gamma})$. Therefore

$$\varphi_i^0(C_{\gamma}, \beta_1, \dots, \beta_l, S, \sigma_2^*) = (\beta_1, \dots, \beta_{i-1}, \gamma, \beta_{i+1}, \dots, \beta_l).$$

2. Let $\varphi_i^0(C_{\gamma}, \beta_1, \dots, \beta_l, S, \sigma_2^*) = (\beta_1, \dots, \beta_{i-1}, \Delta, \beta_{i+1}, \dots, \beta_l)$. By the definition of φ_i^0 , in $M_{s_2}(C_{\gamma})$ there exist sets σ_1, σ_2 such that $P_i(C_{\gamma}, \sigma_1) \neq P_i(C_{\gamma}, \sigma_2)$. But (2) implies that σ_1, σ_2 are contained in $M_{s_1}(C_{\gamma})$. Therefore, $\varphi_i^0(C_{\gamma}, \alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_l, S, \sigma_1^*) = (\alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_l)$.

Paragraphs 1 and 2 imply the monotonicity of φ_i^0 , $i=1, 2, \dots, l$.

We show that $A = \{A_{\pi^*}, \varphi_1^0, \dots, \varphi_l^0, S\}$ is a majorant algorithm. Let $B \in \pi$, $B = \{A_{\pi^*}, \varphi_1, \dots, \varphi_l, S\}$, and let $\alpha_i = \Delta$, $\varphi_i(C_{\gamma}, \alpha_1, \dots, \alpha_l, S, \sigma^*) = (\alpha_1, \dots, \alpha_{i-1}, \gamma, \alpha_{i+1}, \dots, \alpha_l)$, $\gamma \in \{0, 1\}$. Then $\varphi_i^0(C_{\gamma}, \alpha_1, \dots, \alpha_l, S, \sigma^*) = (\alpha_1, \dots, \alpha_{i-1}, \beta, \alpha_{i+1}, \dots, \alpha_l)$. We assume that $\beta \neq \gamma$. If $\beta \in \{0, 1\}$, then $\beta = \gamma$, since otherwise either φ_i or φ_i^0 transforms the information set σ^* into a set which is not an information set.

Let $\beta = \Delta$. Then it follows from the definition of φ_i^0 that the set $M_{\gamma}(C_{\gamma})$ contains σ_1 and σ_2 such that $P_i(C_{\gamma}, \sigma_1) \neq P_i(C_{\gamma}, \sigma_2)$.

We isolate in $M_{\gamma}(C_{\gamma}^1)$ elements σ_1^* and σ_2^* such that $\sigma_1^* \in I(\sigma_1)$, $\sigma_2^* \in I(\sigma_2)$. It is obvious that $S(C_{\gamma}^1, \sigma_1^*) \approx S(C_{\gamma}^1, \sigma_2^*)$. Therefore $\varphi_i(C_{\gamma}, \alpha_1, \dots, \alpha_l, S, \sigma_1^*) = \varphi_i(C_{\gamma}, \alpha_1, \dots, \alpha_l, S, \sigma_2^*) = \varphi_i(C_{\gamma}, \alpha_1, \dots, \alpha_l, S, \sigma^*)$.

From the elements σ_1^*, σ_2^* we select that for which $\gamma \neq P_i(C_{\gamma}, \sigma_j)$, $\sigma_j \in \{\sigma_1^*, \sigma_2^*\}$. Let such an element be σ_1^* . Then the application of φ_i and the replacement in σ_1^* of the element C_{γ}^1 by the element $(C_{\gamma}^1)' = C_{\gamma}^{\alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_l}$ $(X^{\gamma}, Y^{\gamma}, Z^{\gamma})$ takes σ_1^* out of $I(\sigma_1)$, therefore $\gamma = \Delta$. The theorem is proved.

Let $\{\sigma\}$ be a family of sets of subschemes. We suppose that for all (C_{γ}, σ) , $C_{\gamma} \in \sigma$, $\sigma \in \{\sigma\}$ distinct neighborhoods $S^1(C_{\gamma}, \sigma)$, $S^2(C_{\gamma}, \sigma)$, have been introduced, where $S^1(C_{\gamma}, \sigma) \in \{\sigma\}$, $S^2(C_{\gamma}, \sigma) \in \{\sigma\}$. Let A_1 and A_2 be algorithms which are majorant for π_{σ_1} and π_{σ_2} respectively.

Theorem 3.

Let $S^1(C_{\gamma}, \sigma) \subseteq S^2(C_{\gamma}, \sigma)$, then $A_1 \leq A_2$.

Proof. The theorem will be proved if the validity of the following proposition is established.

Let $A_1 \in A(\varphi_{11}, \dots, \varphi_{l1}, P_1, \dots, P_l, S)$, $A_2 \in A(\varphi_{21}, \dots, \varphi_{2l}, P_1, \dots, P_l, S)$, $\varphi_{2i}(C_{\gamma}, \alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_l, S^2, \sigma^*) = (\alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_l)$. Then $\varphi_{1i}(C_{\gamma}, \alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_l, S^1, \sigma^*) = (\alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_l)$.

By the definition of φ_{2i} , in the set $M_{\sigma^2}(C_{\gamma})$ there exist elements σ_1 and σ_2 such that $P_i(C_{\gamma}, \sigma_1) = 0$, $P_i(C_{\gamma}, \sigma_2) = 1$. We show that $\sigma_1 \in M_{\sigma^1}(C_{\gamma})$ and $\sigma_2 \in M_{\sigma^1}(C_{\gamma})$, that is, $S^1(C_{\gamma}, \sigma_1) = S^2(C_{\gamma}, \sigma_2)$. Let $\sigma^* \in I(\sigma)$. We have $S^2(C_{\gamma}, \sigma) \subseteq \sigma_1$, $S^2(C_{\gamma}, \sigma) \subseteq \sigma_2$, $S^2(C_{\gamma}, \sigma_1) = S^2(C_{\gamma}, \sigma) = S^2(C_{\gamma}, \sigma_2)$, $S^1(C_{\gamma}, \sigma_1) \subseteq S^2(C_{\gamma}, \sigma_1)$, $S^1(C_{\gamma}, \sigma_2) \subseteq S^2(C_{\gamma}, \sigma_2)$. Therefore $S^1(C_{\gamma}, \sigma_1) \subseteq S^2(C_{\gamma}, \sigma) \subseteq \sigma_1$, $S^1(C_{\gamma}, \sigma_2) \subseteq S^2(C_{\gamma}, \sigma) \subseteq \sigma_2$.

From part 3 of the definition of neighborhood (see [1]) it follows that $S^1(C_{\gamma}, \sigma_1) = S^1(C_{\gamma}, S_2(C_{\gamma}, \sigma))$, $S^1(C_{\gamma}, \sigma_2) = S^1(C_{\gamma}, S^2(C_{\gamma}, \sigma))$.

By hypothesis of the theorem $S^2(C_{\gamma}, \sigma) \in (\sigma, S^1(C_{\gamma}, S^2(C_{\gamma}, \sigma)))$ is defined, therefore $S^1(C_{\gamma}, \sigma_1) = S^1(C_{\gamma}, \sigma_2)$.

In $M_{\sigma^1}((C_{\gamma}^1)')$ there exist sets σ_1^* and σ_2^* such that $\sigma_1^* \in I(\sigma_1)$, $\sigma_2^* \in I(\sigma_2)$, that is $S^2((C_{\gamma}^1)', \sigma_1^*) \approx S^2((C_{\gamma}^1)', \sigma_2^*)$. Therefore $S^1((C_{\gamma}^1)', \sigma_1^*) \approx S^1((C_{\gamma}^1)', \sigma_2^*)$. The last

equation and the condition $P_i(C_7, \sigma_1) \neq P_i(C_7, \sigma_2)$ imply that $\varphi_i^i(C_7, \alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_i, S^1, \sigma^*) = (\alpha_1, \dots, \alpha_{i-1}, \Delta, \alpha_{i+1}, \dots, \alpha_i)$.

The theorem is proved.

Translated by J. Berry

REFERENCES

1. ZHURABLEV, Yu. I. Local algorithms for the calculation of information. I. *Kibernetika*, No. 1, 12-20, 1965.
2. YABLONSKII, S. V. Fundamental concepts of cybernetics. In: *Problems of cybernetics* (Probl. kibernetiki), 7-39, No. 2, Fizmatgiz, Moscow, 1959.

AN APPROACH TO THE CONSTRUCTION OF OPTIMAL RECOGNITION ALGORITHMS FOR LARGE CONTROL TABLES*

A. G. TSERKOVNYI

Moscow

(Received 4 May 1975)

A RECOGNITION algorithm from the class of estimate calculation algorithms is described. Approximate formulas for calculating the parameter values of the algorithm are presented. A scheme of a method of optimizing this algorithm for working with large training and control tables is presented.

1. Introduction

The algorithm described belongs to the class of algorithms for the calculation of estimates (voting algorithms) [1]. Its principal features may be regarded as: orientation to the solution of comparatively large-scale problems; the possibility of using features from different alphabets in the initial descriptions; partial, as distinct from standard voting algorithms, and optimization carried out with respect to the parameters of the algorithm. As a rule, in known algorithms for the calculation of estimates optimization of the functional of quality of recognition is performed on the parameter space of the algorithm. In this case because of the large dimension of the problem it is impossible to use this approach. Therefore the following process of optimization of the algorithm is proposed. First the approximate values of the parameters of the algorithm are calculated and recognition is performed with these values. At the next stage, using the results of the approximate recognition, the parameters of the algorithm are varied by some amount and a study is made of the effect of the variability of specific parameters on the recognition quality. Then a group of parameters is selected whose variation most substantially affects the improvement of the quality recognition. Then the algorithm is optimized on only the selected subspace of parameters.

2. Description of the class of algorithms

As usual there are two sets of objects, combined in the tables $T_{n,m,t}^1$ and $T_{n,m,t}^2$, called the standard and the control table respectively. The subscripts in the notation $T_{n,m,t}$ have the following meaning: n is the number of features specifying the description of the object to

be recognized, m is the number of objects in the corresponding table, l is the number of classes to which the objects in the table are assigned.

The recognition algorithm is specified by four groups of parameters:

- 1) the parameters p_1, \dots, p_n — the information weights of the features (columns of the table $T_{n,m,l}$);
- 2) the parameters $\gamma_1, \dots, \gamma_m$ — the information weights of the objects (rows of the table $T_{n,m,l}$);

(The information weight has the meaning defined in [2].)

- 3) the parameters $\epsilon_1, \dots, \epsilon_n$ — the thresholds of accuracy in the comparison of the values of the features;

- 4) the parameter λ — the threshold value of the decision rule.

We assume that we are given the values of all the parameters of the algorithm: $p_1, \dots, p_n, \gamma_1, \dots, \gamma_m, \epsilon_1, \dots, \epsilon_n, \lambda$. We describe the procedure of recognition of some object S (row S of $T_{n,m,l}^2$), that is, its assignment to one of the classes $\{K_0, K_1, \dots, K_l\}$ (if the object is assigned to the class K_0 , then this means that the algorithm fails to recognize the given object). The object S of the control set is assigned to one of the classes by sequential comparison with all the objects of the standard set and by calculation of a definite family of estimates.

Therefore, let us compare an object S of the control set and an object S_t of the standard set. The descriptions of these objects are specified by the following values of the features: $S = \{\alpha_1, \dots, \alpha_n\}$ and $S_t = \{\beta_1, \dots, \beta_n\}$.

We will consider that the values of the corresponding features are identical if

$$\rho(\alpha_i, \beta_i) \leq \epsilon_i,$$

where $\rho(\alpha, \beta)$ is a numerical function satisfying the conditions $\rho(\alpha, \beta) = \rho(\beta, \alpha)$, $\rho(\alpha, \alpha) = 0$, $\rho(\alpha, \beta) > 0$, if $\alpha \neq \beta$.

We assume that for the specified objects (the rows S and S_t) the features with numbers i_1, \dots, i_v were identical (in the sense defined above). The proximity function of these rows can be determined in the form $r(S, S_t) = n - v$.

We then construct estimates similar in meaning to the estimates introduced in the standard voting procedures [1].

1. The row by row estimate:

$$\Gamma(S, S_t) = \gamma_t(p_{i_1} + \dots + p_{i_v}),$$

where γ_t is the information weight of the row S_t , and p_{i_1}, \dots, p_{i_v} are the information weights of the features whose values are identical in the rows S and S_t .

2. The estimate of the row S for the class K_j :

$$\Gamma_j(S) = \frac{1}{m_j - m_{j-1}} \sum_{S_i \in K_j} \Gamma(S, S_i), \quad j=1, 2, \dots, l,$$

where $m_j - m_{j-1}$ is the number of rows in the class K_j .

3. The decision rule. After the estimates $\Gamma_1(S), \dots, \Gamma_l(S)$ of the row S have been calculated for each of the classes, we use the decision F given in the following form:

$$F[\Gamma_1(S), \dots, \Gamma_l(S)] = \begin{cases} u, & \text{if } \Gamma_u(S) - \Gamma_j(S) \geq \lambda, u \neq j, j=1, 2, \dots, l, \\ 0 & \text{otherwise} \end{cases}$$

Therefore, we assign the row S of the control table to one of the classes $\{K_0, \dots, K_l\}$. Calculating the corresponding estimates and applying the decision rule successively for each of the objects of the control table, we perform the recognition of all the objects of the table. To estimate the recognition performed it is necessary to introduce a functional of quality of recognition with given parameter values, which can be written in the following form:

$$\varphi = m^*/m',$$

where m^* is the number of correctly recognized objects, and m' is the total number of objects in the control table.

By varying the specific values of the parameters of the algorithm we obtain different values of the quality functional φ . Below we describe the calculation of the approximate values of the parameters of the algorithm, with which the first stage of the recognition procedure is performed.

3. Approximate formulas for selecting the parameters of the algorithm

1. The calculation of the parameters ϵ_i . We first note that as $\rho(\alpha, \beta)$ for comparing the values of the features we can take the simple function $\rho(\alpha, \beta) = |\alpha - \beta|$, but in this case the scale of values of the features may require recoding.

Therefore, we consider a procedure for calculating the accuracy threshold ϵ_i for the feature number i . Let the feature i be able to assume one of the values $\{d_1, \dots, d_s\}$. We extract from the tables $T_{n,m,i}^1$ and $T_{n,m',i}^2$ only the i -th columns, that is, only the values assumed by the i -th features in the objects described. We will consider that we are given the tables $t_{m,i}^1$ and $t_{m',i}^2$, in which the descriptions of the objects are determined by only the one feature i , and we carry out the procedure of recognition of one column from another with different values of the parameter ϵ_i . It is obvious that as possible values of ϵ_i it is advisable to take only $\epsilon_i = d_0, \dots, d_s$, where $d_0 = 0$. As a result of the calculations performed we obtain a matrix $A = \|a_{ij}\|_{(s+1) \times s}$, where the rows correspond to different values of ϵ_i , and the columns to different possible values of the feature i , which are recognized with the given ϵ_i . The element a_{ij} of the matrix A represents the number of correctly recognized values d_j of the feature for some $\epsilon = \epsilon_i$. It should be noted that when the conditions

$$d_j \leq d_0 + \epsilon_i, \quad d_j \geq d_s - \epsilon_i$$

are satisfied, the element $a_{ij} = 0$, since the value of ϵ_i is so great that it spans the whole possible scale of values for the feature i , and in this case correct recognition for d_j is impossible. Therefore, for the

calculation of the best value of ϵ_i it is necessary to consider not all the $s \times (s + 1)$ elements of the matrix A , but rather fewer.

Summing the values of the elements of the matrix A along the rows, we obtain the total number of correctly recognized possible values of the given feature. As an approximate ϵ_i we choose that which corresponds to the maximum number of correctly recognized values of the feature i . If there are several such ϵ_i , then we choose the greatest of the ϵ_i , since in this case for the same quality of recognition we specify less strict conditions on the coincidence of various possible values of the feature i .

2. Calculation of the parameters p_i . After determining the best ϵ_i we find p_i — the information weight of the feature i , also using the matrix A , by the following formula:

$$p_i = q_i / m' \log_2 s,$$

where q_i is the maximum number of correctly recognized values of the feature i , otherwise, the sum of the row of the matrix A corresponding to the best ϵ_i ; s is the number of possible values of the scale of the feature i ; m' is the number of objects in the control table. (In this formula \log_2 is used, since we consider the information weight on 1 information bit.)

3. Calculation of the parameters γ_j . Having the values of the parameters ϵ_i and p_i , we put $\gamma_j = 1$, $j = 1, 2, \dots, m$, and perform the recognition procedure with these parameters. Then the γ_j can be calculated by the following formula:

$$\gamma_j = \left[\sum_{i \in K_j} \Gamma(S_i, S_i') \right]^{-1} \sum_{i \in K_j} \Gamma(S_i, S_i');$$

here the row S_j belongs to the class K_j . This definition of γ_j treats the information weight of the object S_j as the ratio of the estimate of the "proximity" of the object to its class to the estimate of the "proximity" to all the remaining classes.

4. Calculation of the parameter λ . We perform the recognition procedure with the values of ϵ_i , p_i and γ_j previously determined. We specify the decision rule as follows:

$$F[\Gamma_1(S), \dots, \Gamma_l(S)] = \begin{cases} u, & \text{if } \Gamma_u(S) - \Gamma_j(S) > 0, u \neq j, j = 1, 2, \dots, l, \\ 0 & \text{otherwise.} \end{cases}$$

We will simultaneously calculate the quantity

$$\lambda_r = \Gamma_u(S_r) - \max_{j \neq u} \Gamma_j(S_r).$$

As the best value of λ we choose $\lambda = \min_r \lambda_r$, where r runs through only the correctly recognized objects.

Having determined by the procedure described approximate values of the parameters of the algorithm p_1, \dots, p_n , $\gamma_1, \dots, \gamma_m$, $\epsilon_{i1}, \dots, \epsilon_{iq}$, λ , we construct the recognition of the table $T_{n,m',l}^2$ by the table $T_{n,m,l}^1$ with these values of the parameters. To this recognition there corresponds a definite value of the quality functional. The next problem will consist of such a variation of the parameters as will lead to an improvement of the quality of recognition.

4. Necessary conditions for the values of the variation of parameters

Because of the great dimension of the problem it is difficult to find an algorithm extremal with respect to all the parameters.

Therefore a method is proposed enabling us to find a subset of parameters, whose variation has the most substantial effect on the improvement of the quality of recognition. The idea is as follows: varying the parameters subject to the condition that this variation of them does not worsen the original recognition, we ascertain which the parameters are whose variation most substantially affects the improvement of the recognition of the objects at the first stage.

Therefore, we take one of the classes of objects, let us say the class K_l . The objects of this class (S) after the first stage are divided into two subsets: correctly recognized, that is, those for which

$$\Gamma_i(S) - \max_{j \neq i} \Gamma_j(S) \geq \lambda,$$

and incorrectly recognized, which include also objects which the algorithm has refused to recognize.

We vary the values of the information weights p_i by some quantity Δp_i and that of the information weights of the objects γ_j by $\Delta \gamma_j$.

Therefore, the previously introduced value $\Gamma(S, S_j)$ of the estimate of the row S for the row S_j changes its value to $\Gamma'(S, S_j)$:

$$\begin{aligned} \Gamma'(S, S_j) &= (\gamma_j + \Delta \gamma_j) (p_{i_1}^j + \Delta p_{i_1}^j + \dots + p_{i_k}^j + \Delta p_{i_k}^j) \\ &= \gamma_j (p_{i_1}^j + \dots + p_{i_k}^j) + \Delta \gamma_j (p_{i_1}^j + \dots + p_{i_k}^j) + \gamma_j (\Delta p_{i_1}^j + \dots + \Delta p_{i_k}^j) \\ &\quad + \Delta \gamma_j (\Delta p_{i_1}^j + \dots + \Delta p_{i_k}^j) = \Gamma(S, S_j) + \Delta \gamma_j (p_{i_1}^j + \dots + p_{i_k}^j) \\ &\quad + \gamma_j (\Delta p_{i_1}^j + \dots + \Delta p_{i_k}^j) + o(\Delta), \end{aligned} \quad (1)$$

where γ_j is the information weight of the row S_j , $p_{i_1}^j, \dots, p_{i_k}^j$ are the information weights of the features for which the rows S and S_j agree.

Neglecting the last term in (1), we obtain a transformation $\pi_j(S)$ for estimating the row S after the row S_j :

$$\begin{aligned} \pi_j(S) &= \Gamma'(S, S_j) - \Gamma(S, S_j) = \Delta \gamma_j (p_{i_1}^j + \dots + p_{i_k}^j) \\ &\quad + \gamma_j (\Delta p_{i_1}^j + \dots + \Delta p_{i_k}^j). \end{aligned}$$

In the class K_l let the objects $1, 2, \dots, r$ have been recognized correctly, and the objects $r+1, \dots, m_l$ incorrectly. We have varied the values of the parameters p_i and γ_j . It is obviously necessary that the correctly recognized objects with the numbers $1, 2, \dots, r$ should be correctly recognized with the new parameter values also. Therefore, a condition of the form

$$\max(\Gamma'_i, \Gamma_j) - \min(\Gamma'_i, \Gamma_j) \leq \lambda,$$

must be satisfied, where j runs through the correctly recognized rows. Using this necessary condition, we construct a system of inequalities for the increments of the estimates for the class K , for the correctly recognized rows:

$$\frac{1}{m_t} \sum_{j=1}^{m_t} \pi_j(S_i) \leq \lambda,$$

.

$$\frac{1}{m_t} \sum_{j=1}^{m_t} \pi_j(S_r) \leq \lambda,$$

or in expanded form

[illegible]

where m_r is the number of objects in the class K_r .

We have obtained a system of linear inequalities in Δp_i and $\Delta \gamma_j$. After solving system (2), we obtain some constraints imposed on the Δp_i , $\Delta \gamma_j$. These constraints specify the domain of possible variations of the parameters p_i and γ_j . It should be noted that on solving a system of the type (2) for only one class K_i , we obtain constraints possibly not on all the parameters p_i , but only on the parameters γ_j relating to objects occurring in the class K_i . By constructing systems of linear inequalities (2) for the sets of correctly recognized objects of all the classes K_1, \dots, K_l and uniting the resulting domains of variation of the parameters, we find some common domain of possible variations of the parameters of the algorithm p_i and γ_j , characterized by the fact that the variation of the parameters in this domain, in any case, does not worsen the already attained recognition quality. The further aim is that the variation of the parameters in this domain should improve the recognition of the objects which at the first stage were not referred to their proper class.

We write down formally the conditions necessary for the correct recognition of each of the objects not previously recognized. These conditions consist of the satisfaction of the decision rule for each of the objects with the new values of the parameters of the algorithm (that is, the estimate for its own class must be greater than the estimates for any other class taking into account the threshold value λ). Returning once more to the class K_r , we construct a system of inequalities (3) for the first of the incorrectly recognized objects, whose number is $r + 1$:

[illegible]

or in expanded form

$$\begin{aligned} & \frac{1}{m_i} \sum_{j=1}^{m_i} [\Delta\gamma_j(p_{i1}^j + \dots + p_{i, r+1}^j) + \gamma_j(\Delta p_{i1}^j + \dots + \Delta p_{i, r+1}^j)] \\ & - \frac{1}{m_q} \sum_{j=1}^{m_q} [\Delta\gamma_j(p_{q1}^j + \dots + p_{q, r+1}^j) + \gamma_j(\Delta p_{q1}^j + \dots + \Delta p_{q, r+1}^j)] > \Gamma_q - \Gamma_i + \lambda, \end{aligned} \quad (4)$$

$q=1, 2, \dots, l.$

where m_j is the number of objects in the class K_j , Γ_j are the row estimates obtained for the corresponding class K_j at the first stage of recognition.

Similar systems must be constructed for each of the previously incorrectly recognized objects. We check each of the systems of linear inequalities of the form (4) for consistency. Objects to which correspond consistent systems of inequalities constitute a set of rows for which correct recognition is "potentially" possible. We solve each of the consistent systems (4) with constraints on the unknowns Δp_i and $\Delta\gamma_j$ obtained from the solutions of the systems of inequalities (2). In each case, if for some value of Δp_i or $\Delta\gamma_j$ there is a non-zero solution satisfying the constraints, the corresponding parameter p_i or γ_j receives some mark.

As a result to each of the parameters of the recognition algorithm p_i , γ_j there corresponds a certain number of marks, which indicates how many times a non-zero variation of this parameter has given an improved recognition quality. Therefore, in the conditions of the multidimensional problem, we must recognize that the variation of parameters most efficient from the point of view of the improvement of recognition, is that which has obtained the greatest number of marks.

Translated by J. Berry.

REFERENCES

1. ZHURAVLEV, Yu. I. and NIKIFOROV, V. V. Recognition algorithms based on the calculation of estimates. *Kibernetika*, No. 3, 1-11, 1971.
2. ZHURAVLEV, Yu. I., KAMILOV, M. M. and TULYAGANOV, Sh. E. *Algorithms for the calculation of estimates and their application* (Algoritmy vychisleniya otsenok i ikh primeneniye), "Fan", Tashkent, 1974.

SHORT COMMUNICATIONS

THE CONVERGENCE OF MONOTONIC ITERATIVE PROCESSES*

A. Yu. OSTROVSKII

Moscow

(Received 7 July 1975; revised 3 May 1976)

ESTIMATES are given of the convergence of monotonic iterative processes intended for solving problems of the form $x = A_1x - A_2x + f$ subject to the conditions $A_1 > 0$, $A_2 > 0$, $\|A_1 + A_2\| < 1$.

Many technical problems reduce to solving systems of linear algebraic equations of the form

$$x = Ax + f, \quad (1)$$

where A is a matrix, all of whose coefficients a_{ij} are strictly positive, and $\|A\| < 1$. Such systems are often solved by using the classical process of successive approximation

$$x_{k+1} = Ax_k + f.$$

Here if the norm of the matrix A is close to unity, as is often the case in applied problems, then this process converges so slowly that it cannot be regarded as really suitable for calculations. However, if the spread of the coefficients of the matrix is comparatively small, then there exists for the solution of problem (1) an efficient method of accelerating the convergence, proposed in [1] (see also [2]).

In this paper the rate of convergence of this method is estimated. The estimates are given in terms of the so-called θ -norm of the matrix A , introduced and investigated in [3].

1. We consider Eq. (1), where A is a positive matrix with norm less than unity. The iterative process proposed in [1] is constructed as follows. Initial approximations u_0 and v_0 satisfying the relations

$$u_0 \leq v_0, \quad u_0 \leq Au_0 + f, \quad Av_0 + f \leq v_0$$

are chosen (here and below the notation $u \leq v$ means that the components of the vector $v - u$ are non-negative). A number of simple methods exist for finding such u_0, v_0 . In particular, if

$$\|A\| = \max_i \sum_{j=1}^n a_{ij},$$

we can take

$$u_0 = -sz, \quad v_0 = tz, \quad z = (1, 1, \dots, 1),$$

where

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 233-238, 1977.

$$s \geq \max_i \left[f_i \left(\sum_{j=1}^n a_{ij} - 1 \right)^{-1} \right], \quad t \geq \max_i \left[f_i \left(1 - \sum_{j=1}^n a_{ij} \right)^{-1} \right].$$

The successive approximations u_k , v_k and u_k^* , v_k^* are found by the following formulas:

$$\begin{aligned} u_0^* &= u_0, & v_0^* &= v_0, \\ u_{k+1} &= A u_k^* + f, & v_{k+1} &= A v_k^* + f, \\ u_{k+1}^* &= \frac{1}{1+p_{k+1}} (u_{k+1} + p_{k+1} v_{k+1}), & v_{k+1}^* &= \frac{1}{1+q_{k+1}} (v_{k+1} + q_{k+1} u_{k+1}), \\ k &= 1, 2, \dots \end{aligned} \quad (2)$$

Here the non-negative parameters p_{k+1} , q_{k+1} are so chosen that at each step the relations

$$u_k^* \leq u_{k+1}, \quad v_{k+1} \leq v_k^*. \quad (3)$$

are satisfied. As is shown in [1], the resulting successive approximations converge to the solution x of problem (1). The process (2) is monotonic:

$$u_{k-1}^* \leq u_k \leq u_k^* \leq x \leq v_k^* \leq v_k \leq v_{k-1}^*. \quad (4)$$

Since the cone of non-negative vectors in R^n is normal [2], (4) implies the inequalities

$$\|u_k^* - x\| \leq c \|u_k^* - v_k^*\|, \quad \|x - v_k^*\| \leq c \|u_k^* - v_k^*\|,$$

where c is the constant of semimonotonicity of the norm. In particular, if $\|x\| = \max |x_i|$, then $c = 1$. Therefore the rate of convergence of the iterative process investigated can be characterized by the rate of decrease of the quantity $\|u_k^* - v_k^*\|$. It will be shown below that for a judicious choice of the parameters p_k , q_k the value of $\|u_k^* - v_k^*\|$ decreases at the rate of a geometric progression.

2. Let x , y be vectors with non-negative components. We define the numbers $\alpha(x, y)$, $\beta(x, y)$ by the equations

$$\alpha(x, y) = \max \{ \alpha : \alpha \geq 0, x \geq \alpha y \}, \quad \beta(x, y) = \max \{ \beta : \beta \geq 0, y \geq \beta x \}$$

and put

$$\theta(x, y) = (\alpha(x, y) \beta(x, y))^{-1}.$$

We note that if x , y are vectors with positive components, then

$$\alpha(x, y) = \min_i (x_i / y_i) > 0, \quad \beta(x, y) = \min_i (y_i / x_i) > 0$$

and therefore $\theta(x, y) < \infty$. A linear positive operator is called [3] a focussing operator, if

$$\theta(Ax, Ay) \leq C^2, \quad C > 0, \quad x \geq 0, \quad y \geq 0, \quad Ax, Ay \neq 0. \quad (5)$$

The least of the numbers C for which (5) is satisfied, is called the θ -norm of the operator A and is denoted by $\theta(A)$.

Every matrix with positive coefficients is a focussing operator with respect to a cone of non-negative vectors in R^n , and

$$\theta(A) = \left(\max_{i \neq j, s \neq t} \frac{a_{is}a_{jt}}{a_{js}a_{it}} \right)^{1/4}.$$

Theorem 1

In the iterative process (2) let the parameters p_k, q_k be defined by the relations

$$\begin{aligned} p_k &= \max \{ p: p \geq 0, u_k - u_{k-1} \geq p(v_{k-1} - v_k) \}, \\ q_k &= \max \{ q: q \geq 0, v_{k-1} - v_k \geq q(u_k - u_{k-1}) \}. \end{aligned} \quad (6)$$

Then the following inequality holds:

$$\|v_{k+1} - u_{k+1}\| \leq \frac{\theta(A) - 1}{\theta(A) + 1} \|A\| \|u_k - v_k\|. \quad (7)$$

Proof. If $u_k^* = u_{k+1}$ ($v_k^* = v_{k+1}$) for some $k = \bar{k}$, then the vector u_k^* (v_k^*) is the solution of problem (1), and the process is complete. Therefore we have to consider that for $k < \bar{k}$

$$u_k^* \neq u_{k+1}, \quad v_{k+1} \neq v_k^*. \quad (8)$$

Formulas (2) imply the equations

$$\begin{aligned} v_{k+1} - u_{k+1} &= \frac{1}{1+q_{k+1}} (A v_k^* + f + q_{k+1} A u_k^* + q_{k+1} f) \\ &\quad - \frac{1}{1+p_{k+1}} (A u_k^* + f + p_{k+1} A v_k^* + p_{k+1} f) \\ &= A \left(\frac{1}{1+q_{k+1}} v_k^* - \frac{1}{1+p_{k+1}} u_k^* \right) - A \left(\frac{p_{k+1}}{1+p_{k+1}} v_k^* - \frac{q_{k+1}}{1+q_{k+1}} u_k^* \right), \end{aligned}$$

whence

$$v_{k+1} - u_{k+1} = \frac{1 - p_{k+1} q_{k+1}}{(1+p_{k+1})(1+q_{k+1})} A (v_k^* - u_k^*).$$

Consequently,

$$\|v_{k+1} - u_{k+1}\| \leq \left| \frac{1 - p_{k+1} q_{k+1}}{(1+p_{k+1})(1+q_{k+1})} \right| \|A\| \|v_k^* - u_k^*\|. \quad (9)$$

We find an upper bound of the numerical factor on the right side of the inequality (9). By the same formulas (2),

$$\begin{aligned} u_{k+1} - u_k^* &= A u_k^* + f - u_k^* \\ &= \frac{1}{1+p_k} A (u_k + p_k v_k) + f - \frac{1}{1+p_k} [A (u_{k-1} + f) + p_k (A v_{k-1} + f)], \end{aligned}$$

whence

$$u_{k+1} - u_k^* = \frac{1}{1+p_k} A [(u_k - u_{k-1}) - p_k (v_{k-1} - v_k)].$$

Similarly,

$$v_h^* - v_{h+1} = \frac{1}{1+q_h} A[(v_{h-1}^* - v_h) - q_h(u_h - u_{h-1}^*)].$$

We put

$$g_h = \frac{1}{1+p_h} [(u_h - u_{h-1}^*) - p_h(v_{h-1}^* - v_h)],$$

$$h_h = \frac{1}{1+q_h} [(v_{h-1}^* - v_h) - q_h(u_h - u_{h-1}^*)].$$

Because of the choice of p_k, q_k the vectors g_k, h_k are non-negative, and the vectors $u_{h+1} - u_h^* = Ag_h, v_h^* - v_{h+1} = Ah_h$ are non-zero by (8). Therefore, by (5)

$$\theta(u_{h+1} - u_h^*, v_h^* - v_{h+1}) \leq \theta^2(A).$$

The equation

$$(p_{h+1}q_{h+1})^{-1} = \theta(u_{h+1} - u_h^*, v_h^* - v_{h+1}),$$

holds by (6), and this implies the relation

$$(p_{h+1}q_{h+1})^{-1} \leq \theta^2(A). \quad (10)$$

We put $(p_{h+1}q_{h+1})^{-1} = v^2$. Then $v \geq 1$, and therefore

$$\left| \frac{1 - p_{h+1}q_{h+1}}{(1+p_{h+1})(1+q_{h+1})} \right| \leq \left| \frac{1 - 1/v^2}{1 + 1/v^2 + 2/v} \right| = \frac{v-1}{v+1}. \quad (11)$$

From (10), (11) and the monotonicity of the function $(x-1)/(x+1)$ there follows the inequality

$$\left| \frac{1 - p_{h+1}q_{h+1}}{(1+p_{h+1})(1+q_{h+1})} \right| \leq \frac{\theta(A) - 1}{\theta(A) + 1},$$

which implies (7). The theorem is proved.

3. Theorem 1 shows that the process (2) is especially efficient in those cases where the norm of the matrix A is close to unity, and the spread of its coefficients is not very great. However, its use is also useful whenever the norm of the matrix A is small. The realization of the process (2) on a computer requires twice the number of operations of the usual method of successive approximation. But the process (2) converges like a geometrical progression with ratio $q = (\theta(A) - 1)(\theta(A) + 1)^{-1} \|A\|$. Therefore it can be regarded as more efficient than the usual method of successive approximation if

$$\frac{\theta(A) - 1}{\theta(A) + 1} \|A\| \leq \|A\|^2,$$

that is,

$$\theta(A) \leq (1 + \|A\|)(1 - \|A\|)^{-1}.$$

The satisfaction of the last inequality is easy to check for specific systems. For the estimation of $\theta(A)$ is convenient to use one of the four relations:

$$\theta(A) \leq \max_i \gamma_i, \quad \theta(A) \leq \max_{i \neq j} (\gamma_i \gamma_j)^{1/2},$$

$$\theta(A) \leq \max_j v_j, \quad \theta(A) \leq \max_{i \neq j} (v_i v_j)^{1/2},$$

where

$$\gamma_i = \max_{s,t} \frac{a_{is}}{a_{it}}, \quad \nu_j = \max_{s,t} \frac{a_{sj}}{a_{tj}}, \quad i, j=1, 2, \dots, n.$$

4. An iterative process similar to the one described can be used to solve the problem

$$x = Ax - Bx + f, \quad (12)$$

where A, B are positive matrices satisfying the condition $\|A+B\| < 1$. In this case the initial approximations u_0, v_0 are so chosen that the relations

$$u_0 \leq v_0, \quad u_0 \leq Au_0 - Bv_0 + f, \quad Av_0 - Bu_0 + f \leq v_0.$$

are satisfied. The successive approximations u_k, v_k and u_k^*, v_k^* are sought by the formulas

$$\begin{aligned} u_0^* &= u_0, & v_0^* &= v_0, \\ u_{k+1} &= Au_k^* - Bv_k^* + f, & v_{k+1} &= Av_k^* - Bu_k^* + f, \\ u_{k+1}^* &= \frac{1}{1+t_{k+1}} (u_{k+1} + t_{k+1}v_{k+1}), & v_{k+1}^* &= \frac{1}{1+t_{k+1}} (v_{k+1} + t_{k+1}u_{k+1}), \\ k &= 1, 2, \dots, \end{aligned} \quad (13)$$

where the parameter $t_k \geq 0$ is so chosen that relations (3) are satisfied. As above, the successive approximations converge monotonically to the solution x of problem (12) [1]. It will be shown below that if the choice of t_k is subject to some natural condition, then the quantity $\|u_k^* - v_k^*\|$ decreases at the rate of a geometrical progression.

Let x, y be vectors with non-negative components. We determine the number $T(x, y)$ by the equation

$$T(x, y) = \max\{t: t \geq 0, x \geq ty, y \geq tx\}.$$

Obviously,

$$T(x, y) = \min[\min_i (x_i/y_i), \min_i (y_i/x_i)] \leq 1, \quad (14)$$

and if x, y are vectors with positive components, then $T(x, y) > 0$.

Lemma. Let A, B be positive matrices, and let $x = (x_1, \dots, x_n), y = (y_1, \dots, y_n)$ be arbitrary non-negative vectors, at least one of which is non-zero. Then

$$T(Ax + By, Ay + Bx) \geq d > 0, \quad (15)$$

where

$$d = \min[\min_{i,j} (a_{ij}/b_{ij}), \min_{i,j} (b_{ij}/a_{ij})]. \quad (16)$$

Proof. We put

$$\begin{aligned} c_i &= (a_{i1}, \dots, a_{in}, b_{i1}, \dots, b_{in}), & d_i &= (b_{i1}, \dots, b_{in}, a_{i1}, \dots, a_{in}), & i &= 1, 2, \dots, n, \\ z &= (x_1, \dots, x_n, y_1, \dots, y_n). \end{aligned}$$

By (14)

$$\begin{aligned} T(Ax + By, Ay + Bx) &= \min \left[\min_i \frac{(c_i, z)}{(d_i, z)}, \min_i \frac{(d_i, z)}{(c_i, z)} \right] \\ &\geq \min \left[\min_i \min_{z \geq 0} \frac{(c_i, z)}{(d_i, z)}, \min_i \min_{z \geq 0} \frac{(d_i, z)}{(c_i, z)} \right]. \end{aligned}$$

It is convenient to denote the components of the vectors c_i, d_i by $c_{ij}, d_{ij}, j=1, 2, \dots, 2n$, respectively. Obviously,

$$\min_{z \geq 0} [(c_i, z)/(d_i, z)] = \min \{ \lambda : \lambda = (c_i, z), z \geq 0, (d_i, z) = 1 \}.$$

Therefore

$$\min_{z \geq 0} [(c_i, z)/(d_i, z)] = \min_j (c_{ij}/d_{ij}), \quad j=1, 2, \dots, 2n.$$

Similarly,

$$\min_{z \geq 0} [(d_i, z)/(c_i, z)] = \min_j (d_{ij}/c_{ij}), \quad j=1, 2, \dots, 2n.$$

The last two equations imply the validity of (15). The lemma is proved.

Theorem 2

In the iterative process (13) let the parameters t_k be defined by the equation

$$t_k = \max \{ t : t \geq 0, u_k - u_{k-1} \geq t(v_{k-1} - v_k), v_{k-1} - v_k \geq t_k(u_k - u_{k-1}) \}. \quad (17)$$

Then the inequality

$$\|v_{k+1} - u_{k+1}\| \leq \frac{S(A, B) - 1}{S(A, B) + 1} \|A + B\| \|v_k^* - u_k^*\|, \quad (18)$$

holds, where

$$S(A, B) = \max[\max_{i,j} (a_{ij}/b_{ij}), \max_{i,j} (b_{ij}/a_{ij})].$$

Proof. Reasoning as in the proof of Theorem 1, we will assume that for $k < \bar{k}$

$$u_k^* \neq u_{k+1}, \quad v_{k+1} \neq v_k^*. \quad (19)$$

Formulas (13) imply the equations

$$\begin{aligned} v_{k+1} - u_{k+1} &= (A+B)(v_k^* - u_k^*), \\ v_{k+1} - u_{k+1} &= \frac{1-t_{k+1}}{1+t_{k+1}} (v_{k+1} - u_{k+1}), \\ v_{k+1} - u_{k+1} &= \frac{1-t_{k+1}}{1+t_{k+1}} (A+B)(v_k^* - u_k^*). \end{aligned}$$

Consequently,

$$\|v_{k+1} - u_{k+1}\| \leq \left| \frac{1-t_{k+1}}{1+t_{k+1}} \right| \|A+B\| \|v_k^* - u_k^*\|. \quad (20)$$

Using formulas (13) we write the equations

$$u_{h+1} - u_h^* = A g_h + B h_h, \quad v_h^* - v_{h+1} = A h_h + B g_h,$$

where

$$g_h = \frac{1}{1+t_h} [(u_h - u_{h-1}^*) - t_h (v_{h-1}^* - v_h)],$$

$$h_h = \frac{1}{1+t_h} [(v_{h-1}^* - v_h) - t_h (u_h - u_{h-1}^*)].$$

Because of the choice of t_k the vectors g_k , h_k are non-negative, and by (19), $u_{h+1} - u_h^* \neq 0$, $v_h^* - v_{h+1} \neq 0$. Therefore, the lemma implies the estimate

$$T(u_{h+1} - u_h^*, v_h^* - v_{h+1}) \geq d, \quad (21)$$

where d is defined by formula (16). In accordance with (17) we have

$$t_{h+1} = T(u_{h+1} - u_h^*, v_h^* - v_{h+1}),$$

therefore (20), (21) imply the estimate

$$\|v_{h+1}^* - u_{h+1}^*\| \leq \frac{1-d}{1+d} \|A+B\| \|v_h^* - u_h^*\|.$$

This implies (18), since $S(A, B) = 1/d$.

Translated by J. Berry.

REFERENCES

1. STETSENKO, V. Ya. A method of accelerating the convergence of iterative processes. *Dokl. Akad. Nauk SSSR*, 178, 5, 1021-1024, 1968.
2. KRASNOSEL'SKII, M. A., VAINIKKO, G. M., ZABREIKO, P. P., RUTITSKII, Ya. B. and STETSENKO, V. Ya. *The approximate solution of operator equations* (Priblizhennoe reshenie operatornykh uravnenii), "Nauka", Moscow, 1969.
3. ZABREIKO, P. P., KRASNOSEL'SKII, M. A. and POKORNYI, Yu. V. On a class of linear positive operators. *Funkts. analiz ego prilozh.* 5, 4, 9-17, 1971.

A METHOD OF REGULARIZING THE INVERSE RADON TRANSFORMATION IN A MEDICO-BIOLOGICAL PROBLEM*

N. P. LIPATOV

Moscow

(Received 19 June 1975; revised 2 February 1976)

A REGULARIZING algorithm is constructed for the problem of reconstructing a function from its radon transformation. The two-dimensional case is considered.

Recently in a number of fields of technology and medicine the problem of the reconstruction of a section of a body has become urgent. By the reconstruction of a section we mean the determination of the coefficient of linear absorption of radiation $f(x, y)$ at an arbitrary

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 238-241, 1977.

point of the section. One possible way to solve this problem is as follows. A number of transilluminations of the body investigated are made with a narrow beam of radiation in the plane of the section; by processing the resulting data it is possible to obtain the value of $f(x, y)$ at an arbitrary point.

When a narrow beam of radiation passes through the body investigated its intensity on emergence is defined as follows:

$$I(\alpha, p) = I_0 \exp \left[- \int_{(\alpha, p)} f(x, y) dl \right].$$

Here α and p are the parameters of the straight line along which the transillumination occurs, and I_0 is the intensity of the incident radiation. The symbol (α, p) under the integral sign shows that the integration takes place along the straight line with parameters α and p .

It is usually assumed that p cannot assume negative values. But we will permit negative values of p . It is obvious that the straight lines with parameters $\alpha + \pi, p$ and $\alpha, -p$ coincide.

We denote the integral in the exponent \exp by $F(\alpha, p)$. The expression $F(\alpha, p)$ is called the Radon transform of the function $f(x, y)$:

$$F(\alpha, p) = \int_{(\alpha, p)} f(x, y) dl = -\ln \frac{I(\alpha, p)}{I_0}.$$

We denote the operator of the transition from $f(x, y)$ to its Radon transform by P .

Since $I(\alpha, p)$ is known from measurement, and I_0 is given, the problem reduces to the following: to reconstruct a function from its Radon transform.

Since $F(\alpha + \pi, p) = F(\alpha, -p)$ and $I(\alpha + \pi, p) = I(\alpha, -p)$, the measurements of $I(\alpha, p)$ must be made for $0 \leq \alpha < \pi$.

The body investigated has finite dimensions, hence it is natural to assume that $f(x, y)$ vanishes outside a circle of radius R with centre at the origin of coordinates.

It is known that if $f(x, y)$ is continuous, then it is uniquely reconstructed by $F(\alpha, p)$.

We define $\|f\|$ and $\|F\|$ as follows:

$$\|f\| = \sup |f(x, y)|, \quad \|F\| = \sup |F(\alpha, p)|.$$

It is obvious that in the case of such norms the problem of the reconstruction of f from F is ill-posed, therefore it is advisable to construct a regularizing algorithm.

Using geometrical considerations we can deduce the following formula:

$$\iint_{x, y, z} f(x, y) dx dy = \int_{-\infty}^{\infty} F(\alpha, p) dp - \frac{1}{\pi} \int_0^{\infty} \frac{s}{(s^2 - \delta^2)^{1/2}} \left(\int_0^{2\pi} F(\alpha, s) d\alpha \right) ds, \quad (1)$$

where $K_{0,0,\delta}$ denotes the interior of a circle of radius δ with centre at the origin of coordinates.

If we denote by $K_{x,y,\delta}$ the interior of a circle of radius δ with centre at the point x, y , and by $\bar{R}[F, \delta]$ the mean value of $f(x, y)$ in the circle $K_{x,y,\delta}$, then from (1) we obtain

$$\bar{R}[F, \delta] = \frac{1}{\pi\delta^2} \left\{ \int_{-\infty}^{\infty} F(\alpha, p) dp - \frac{1}{\pi} \int_0^{\infty} \frac{s}{(s^2 - \delta^2)^{1/2}} \left[\int_0^{2\pi} F(\alpha, s + x \cos \alpha + y \sin \alpha) d\alpha \right] ds \right\}. \quad (2)$$

In order to verify that $\bar{R}[F, \delta]$ is a regularizing operator, we can use the following theorem (see [1]).

Theorem

Let A be an operator from V into U , and $\bar{R}[u, \delta]$ an operator from U into V , defined for any element of U and any $\delta > 0$, continuous with respect to u . If for any element $z \in V$

$$\lim_{\delta \rightarrow 0} \bar{R}[Az, \delta] = z,$$

then the operator $\bar{R}[u, \delta]$ is a regularizing operator for the equation $Az = u$.

In our case V is the set of functions continuous in the circle $K_{0,0,R}$ and vanishing outside this circle, U is a set of continuous functions, where $F(\alpha + \pi, p) = F(\alpha, -p)$ and $F(\alpha, p) = 0$ for $p > R$, and $A = P$.

We impose another constraint on the system of functions V : for any $f(x, y)$ belonging to V let the condition

$$|f(x_1, y_1) - f(x_2, y_2)| \leq D((x_1 - x_2)^2 + (y_1 - y_2)^2)^{1/2}, \quad (3)$$

be satisfied, where D is a constant independent of f .

We estimate the accuracy of the specification of the original information (relative to the accuracy of $I(\alpha, p)$ and I_0), which will be sufficient to reconstruct $f(x, y)$ with accuracy ϵ . For this we prove two propositions.

Proposition 1

If $f(x, y)$ satisfies the requirement (3) and $r \leq \epsilon/2D$, then the following inequality is satisfied:

$$\left| f(x, y) - \frac{1}{\pi r^2} \iint_{K_{x,y,r}} f(x, y) dx dy \right| \leq \frac{\epsilon}{2}. \quad (4)$$

The proof is obvious.

Proposition 2

Let $F(\alpha, p)$ and $G(\alpha, p)$ belong to U , $f(x, y)$ belong to V and $F(\alpha, p)$ be the Radon transform of $f(x, y)$, $H(\alpha, p) = F(\alpha, p) - G(\alpha, p)$, $|H(\alpha, p)| \leq \beta$; then the following inequality will be satisfied:

$$\left| \iint_{K_{x,y,\delta}} f(x, y) dx dy - R[G, \delta] \right| \leq \frac{4\beta}{\pi\delta}. \quad (5)$$

Proof. We consider the expression

$$\xi = \frac{1}{\pi\delta^2} \left| \int_{-\infty}^{\infty} H(\alpha, p) dp - \frac{1}{\pi} \int_0^{\infty} \frac{s}{(s^2 - \delta^2)^{1/2}} \left[\int_0^{2\pi} H(\alpha, s + x \cos \alpha + y \sin \alpha) d\alpha \right] ds \right|.$$

Since the first integral under the modular sign is independent of α , it can be integrated with respect to α between the limits 0 and 2π and divided by 2π , without changing its value:

$$\begin{aligned} \xi &= \frac{1}{\pi\delta^2} \left| \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} H(\alpha, s + x \cos \alpha + y \sin \alpha) ds d\alpha \right. \\ &+ \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} H(\alpha + \pi, s + x \cos(\alpha + \pi) + y \sin(\alpha + \pi)) ds d\alpha \\ &\left. - \frac{1}{\pi} \int_0^{\infty} \frac{s}{(s^2 - \delta^2)^{1/2}} \left[\int_0^{2\pi} H(\alpha, s + x \cos \alpha + y \sin \alpha) d\alpha \right] ds \right|. \end{aligned} \quad (6)$$

From (6) it follows that

$$\begin{aligned} \xi &\leq \frac{1}{\pi^2\delta^2} \left| \int_0^{\delta} \int_0^{2\pi} H(\alpha, s + x \cos \alpha + y \sin \alpha) d\alpha ds \right| \\ &+ \frac{1}{\pi^2\delta^2} \left| \int_0^{\infty} \int_0^{2\pi} H(\alpha, s + x \cos \alpha + y \sin \alpha) d\alpha ds \right. \\ &\left. - \int_0^{\infty} \frac{s}{(s^2 - \delta^2)^{1/2}} \int_0^{2\pi} H(\alpha, s + x \cos \alpha + y \sin \alpha) d\alpha ds \right|. \end{aligned}$$

Replacing in the last formula the infinite limits of integration by $2R$ and evaluating the integral, we obtain (5).

It follows from (4) and (5) that to determine the approximate value of $f(x, y)$ with accuracy ϵ it is sufficient that $|F(\alpha, p) - G(\alpha, p)|$ do not exceed $\pi\epsilon^2/16D$. Here G is an approximate value of F , and F is the Radon transform of the function f .

We now determine the relative accuracy of $I(\alpha, p)$ and I_0 . From the inequality

$$\left| \ln \frac{I(\alpha, p)}{I_0} - \ln \frac{J(\alpha, p)}{J_0} \right| \leq \frac{\pi e^2}{16D},$$

where $J(\alpha, p)$ and J_0 are approximate values of $I(\alpha, p)$ and I_0 , it follows that $(\pi e^2/16D)$ is assumed to be a small quantity) if the relative errors of $J(\alpha, p)$ and I_0 do not exceed $\pi e^2/32D$, then the ϵ -accuracy in the determination of $f(x, y)$ will be attained.

In conclusion we note that if on the set of functions $f(x, y)$ we specify the norm L_1 or L_2 and vanishes outside the circle $(K_{0,0,R})$, then the problem (the existence of a solution is assumed) will be incorrect just the same. The operator (2) will be a regularizing operator in this case also.

Translated by J. Berry

REFERENCES

1. TIKHONOV, A. N. and ARSENIN, V. Ya. *Methods of solving ill-posed problems* (Metody resheniya nekorrektnykh zadach), "Nauka", Moscow, 1974.

OPTIMAL QUADRATURE FORMULAS FOR A SPHERE*

V. A. GORDIN

Moscow

(Received 30 May 1975; revised 15 December 1975)

THE PROBLEM of finding the statistically optimal quadrature formula for a sphere is posed. A system of linear algebraic equations satisfied by weights of the quadrature formula is written down. Two examples are given. Asymptotic estimates of the relative error are given.

1. Let the linear functional

$$f(\varphi) = \int_{S^2} \varphi(x) dv(x) \quad (1)$$

be defined in the space of functions on the two-dimensional sphere S^2 . In practice instead of the functional (1) we often consider a functional of the form

$$F(\varphi) = \sum_{j=1}^N a_j \varphi(x_j), \quad \text{where } a_j \in \mathbb{C}, \quad x_j \in S^2. \quad (2)$$

Sometimes the statistics of the functions φ , on which the functionals (1) and (2) act, are known. This situation is encountered, for example, in meteorology [1], where S^2 is a natural object, and the statistics are vast.

In this case the problem of the best choice of the coefficients can be given a probabilistic sense: the functional F best approximates the functional f , if on it there is attained the minimum

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 241-246, 1977.

$$\min_{a_j \in G} M |f(\varphi) - F(\varphi)|^2. \quad (3)$$

The numbers a_j and the points x_j are chosen deterministically for known statistics. (Therefore, the method considered is not of the Monte Carlo type.)

2. Let φ be a homogeneous centred random field on S^2 , that is, for every point $x \in S^2$ we have $M\varphi(x) = 0$ and for any points $x, y \in S^2$ and any element of the group of rotations $g \in SO(3)$ let the correlation function of this field $B(x, y) = M\{\varphi(x)\overline{\varphi(y)}\}$ be invariant with respect to g : $B(x, y) = B(gx, gy)$.

For such a field the following representation holds (see [2, 3]):

$$\varphi(x) = \sum_{|n| \leq l, l=0}^{l=\infty} Y_n^l(x) Z_{ln},$$

where $Y_n^l(x)$ are spherical functions, and Z_{ln} are random quantities, and

$$MZ_{ln} = 0, \quad M\{Z_{ln}, Z_{l_1 n_1}\} = \delta_{ll_1} \delta_{nn_1} \Psi(l) \geq 0.$$

3. We consider the functionals h_{ij} operating by the formula

$$h_{ij} \left(\sum_{|n| \leq l, l=0}^{l=\infty} Y_n^l(x) Z_{ln} \right) = Z_{lj},$$

and their linear shell L . In the linear space L formula (3) specifies the structure of the pre-Hilbert space:

$$\langle g, h \rangle = M\{g(\varphi)\overline{h(\varphi)}\} = \sum_{|n| \leq l, l=0}^{l=\infty} \Psi(l) g(Y_n^l) \overline{h(Y_n^l)}. \quad (4)$$

We denote by H the corresponding Hilbert space. It is finite-dimensional if and only if the measure of $\Psi(l)$ is finite.

4. From the theorem of the perpendicular we deduce the following theorem.

Theorem

Let $g_0, g_1, \dots, g_N \in H$ and g_1, \dots, g_N be linearly independent. In the subspace tight on g_1, \dots, g_N , the best approximation to g_0 in the sense of (4) is the functional $g = \sum_{j=1}^N a_j g_j$ where the numbers a_j satisfy the system

$$\sum_{i=1}^N \langle g_i, g_j \rangle a_i = \langle g_0, g_j \rangle, \quad j=1, 2, \dots, N. \quad (5)$$

In the case where $g_1, \dots, g_N \in H, g_0 \in H$, but $|\langle g_0, g_j \rangle| < \infty$ for all $j \geq 1$, the specification of (3) does not have an exact meaning and the solution of system (5) can be called a "weak" solution of problem (3).

In the case considered in paragraph 1, for the conditions of the theorem to be satisfied it is necessary and sufficient that

$$|B(x, x)| < \infty, \quad \left| \iint_{S^1 \times S^1} B(x, y) dv(x) dv(y) \right| < \infty.$$

Since these conditions are not always satisfied, they represent constraints on the random field φ . For example, for white noise it is impossible to solve the problem of the optimal quadrature formula.

5. The problem of the best approximation of functionals operating on a centred random field is a certain complication of the above. Let $\varphi = \varphi_0 + \varphi_1$, where φ_0 is a deterministic field, and φ_1 is a homogeneous centred random field. Let $|g_j(\varphi_0)| < \infty$ for all $j \geq 0$, g_1, \dots, g_N be linearly independent and $j \geq 1$ exist such that $g_j(\varphi_0) \neq 0$. We will find the minimum (3)

in the class of functionals $g = \sum_{j=1}^N a_j g_j$, giving the unbiased estimate g_0 : $g(\varphi_0) = g_0(\varphi_0)$.

By the method of Lagrange multipliers we obtain a linear system satisfied by the optimal weights a_j in this case:

$$\sum_{i=1}^N \langle g_i, g_j \rangle a_i + \lambda g_j(\varphi_0) = \langle g_0, g_j \rangle, \quad j=1, 2, \dots, N, \quad (6)$$

$$\sum_{i=1}^N a_i g_i(\varphi_0) = g_0(\varphi_0).$$

It is easy to see that with the assumptions made the system (6) is uniquely solvable.

6. Let $g_j = \delta_{x_j}(x)$, $dv(x) = dx$ and $B(x, y) = \exp \{-[\theta(x, y)/a]^2\}$, where $\theta(x, y)$ is the distance along the geodesic between the points $x, y \in S^2$ in radians; a correlation function of this form is characteristic, for example, for geopotential fields [1] for $a = 0.205$. The positive-definiteness of this function, that is, the positiveness of the coefficients of its expansion in Legendre polynomials, has been verified numerically. We consider a latitude-longitude

template $M_1 = \{x_j\}_{j=1}^{j=98}$ with step in latitude $\Delta\theta = 36^\circ$ and in longitude $\Delta\varphi = 15^\circ$. The optimal weights a_j , depending only on $|\theta|$, and the relative error κ of the quadrature formula ($\kappa = \|g - g_0\| \|g_0\|^{-1}$) are given in column A of Table 1.

TABLE 1

$ \theta $, rad.	A	B
1.571	0.130	0.235
0.943	0.055	0.098
0.314	0.089	0.156
κ	0.655	0.861

The result obtained is easy to interpret: the points at latitudes $\pm 54^\circ$ are noticeably correlated with their neighbours, and consequently carry less information than, for example, the polar points which are practically not correlated with other points:

$$a_1 = a_{98} \approx \left(\int_{S^2} B(x, y) dy \right) [B(x, x)]^{-1}.$$

Therefore, the polar points in (2) occur with the greatest weight.

The weights a_j have a similar interpretation for the latitude-longitude template $\mathcal{M}_2 = \{x_j\}_{j=1}^{1202}$ with step in latitude $\Delta\theta = 5.8^\circ$ and in longitude $\Delta\varphi = 9^\circ$, see column A of Table 2.

TABLE 2

$ \theta $	A	B	$ \theta $	A	B
1.571	0.00644	0.00655	0.659	0.01234	0.01249
1.469	0.00154	0.00157	0.558	0.01325	0.01351
1.368	0.00316	0.00322	0.456	0.01401	0.01430
1.267	0.00467	0.00475	0.355	0.01464	0.01492
1.165	0.00616	0.00629	0.254	0.01511	0.01541
1.064	0.00757	0.00772	0.452	0.01543	0.01574
0.963	0.00892	0.00910	0.051	0.01559	0.01590
0.861	0.01016	0.01037	κ	0.140	0.141
0.760	0.01131	0.01155	κ_1		0.160

7. Now let the situation of paragraph 5 hold; $\varphi_0 = 1$, φ_1 the same as in paragraph 6. The weights and the errors for \mathcal{M}_1 and \mathcal{M}_2 are given in column B of Tables 1 and 2 respectively.

As a comparison the relative error κ_1 of the quadrature formula with the template \mathcal{M}_2 and with the weights $a_j = 4\pi/1202$, the same for all points is given.

8. There are a number of papers (see, for example, [4, 5]), in which the problem of finding quadrature formulas for a sphere, exact on the spherical functions $Y_n^l(x)$ for $l \leq M$ is solved. Such a problem is obtained, for example, if we put $\Psi(l) = 1$ for $l \leq M$ and $\Psi(l) = 0$ for $l > M$. Then $\dim H = (M+1)^2$, and, taking $(M+1)^2$ independent functionals g_j , we obtain a basis in H , with respect to which $g_0 \in H$ can be expanded with zero error.

9. The problem of finding the minimum (3) can be generalized, by optimizing not only with respect to the coefficients a_j , but also with respect to the templates $\{x_j\}$. It is then possible to consider both arbitrary templates with a given number of points, and also templates with geometrical constraints, for example, to search for the optimal template in the class of latitude-longitude templates or templates invariant with respect to some group. For $\dim H < \infty$ such problems were solved in [4-6].

10. We consider asymptotic estimates of the error. Let $III = \{x_j\}_{j=1}^N$ and $U_j = \{x \in S^2 \mid \forall i \neq j, \theta(x, x_j) < \theta(x, x_i)\}$. The domains U_j are bounded by a finite number of geometrical arcs and they are convex, and their closures \bar{U}_j cover the sphere:

Let
$$\bigcup_{j=1}^N \bar{U}_j = S^2.$$

$$\rho_j = \max_{x \in U_j} \theta(x, x_j), \quad \rho = \max_{1 \leq j \leq N} \rho_j.$$

Then III is a ρ -network on S^2 .

We represent the functional g_0 as a sum

$$g_0 = \sum_{j=1}^N g^j, \quad \text{where} \quad g^j(\varphi) = \int_{U_j} \varphi(x) dx.$$

We note that by the triangular inequality the estimate

$$\left\| g_0 - \sum_{j=1}^N a_j g_j \right\| = \left\| \sum_{j=1}^N g^j - \sum_{j=1}^N a_j g_j \right\| \leq \sum_{j=1}^N \|g^j - a_j g_j\|,$$

holds, and this implies that

$$\min_{a_j \in \mathbb{C}} \left\| g_0 - \sum_{j=1}^N a_j g_j \right\| \leq \sum_{j=1}^N \min_{a_j \in \mathbb{C}} \|g^j - a_j g_j\|. \quad (7)$$

The minimum on the right side of (7) is found by the theorem of the perpendicular:

$a_j = \langle g^j, g_j \rangle \langle g_j, g_j \rangle^{-1}$, and the error is then

$$\begin{aligned} \min_{a_j \in \mathbb{C}} \|g^j - a_j g_j\| &= (\langle g^j - a_j g_j, g^j - a_j g_j \rangle)^{1/2} = \left\{ \int_{U_j} \int_{U_j} B(x, y) dx dy \right. \\ &\quad \left. - \left[\int_{U_j} B(x, x_j) dx \right]^2 [B(x_j, x_j)]^{-1} \right\}^{1/2}. \end{aligned} \quad (8)$$

We estimate (8) as $\rho \rightarrow 0$. Let the correlation function $B = B(\theta)$ be a Lipschitz function: a $K > 0$ exists such that $|B(\theta_1) - B(\theta_2)| \leq K|\theta_1 - \theta_2|$. In this case we can estimate the integrands in (8):

$$\begin{aligned} |B(x, y) - B(x, x_j)| &= |B(\theta(x, y)) - B(\theta(x, x_j))| \leq K|\theta(x, y) - \theta(x, x_j)| \\ &\leq K\theta(y, x_j), \\ |B(x, x_j) - B(x_j, x_j)| &\leq K|\theta(x, x_j) - \theta(x_j, x_j)| = K\theta(x, x_j), \\ \left| \int_{U_j} \int_{U_j} B(x, y) dx dy - \int_{U_j} B(x, x_j) dx \int_{U_j} dx \right| &\leq K \int_{U_j} dx \int_{U_j} \theta(x, x_j) dx \leq \frac{2K\pi^2 \rho_j^5}{3}, \end{aligned} \quad (9)$$

$$\left| \int_{U_j} B(x, x_j) dx - B(x_j, x_j) \int_{U_j} dx \right| \leq K \int_{U_j} \theta(x, x_j) dx \leq \frac{2K\pi\rho_j^3}{3}, \quad (\text{cont'd})$$

$$\left\{ \iint_{U_j, U_j} B(x, y) dx dy - \left[\int_{U_j} B(x, x_j) dx \right]^2 [B(x_j, x_j)]^{-1} \right\}^{1/2} \leq 2\pi \left[\frac{K\rho^5}{3} \right]^{1/2}.$$

Substituting (9) into (7), we obtain

$$\min_{a_j \in \mathbb{C}} \left\| g_0 - \sum_{j=1}^N a_j g_j \right\| \leq \sum_{j=1}^N \left(\frac{4\pi^2 K \rho_j^5}{3} \right)^{1/2} \leq N \left(\frac{4\pi^2 K}{3} \right)^{1/2} \rho^{5/2}. \quad (10)$$

By [7], there exists a $C > 0$ such that for every $\epsilon > 0$ there exists a $\mathcal{M} = \{x_j\}_{j=1}^N$ such that \mathcal{M} is an ϵ -mesh on S^2 and $N \leq C\epsilon^{-2}$. (In [7] the general result is given. In the case of a two-dimensional sphere we can take as \mathcal{M} a latitude-longitude template. For the estimate of C for S^2 see [8].) Consequently the following theorem is proved.

Theorem

Let $B(\theta)$ be a Lipschitz function with constant K . There exists a template with the number of rows not exceeding N and with an error κ not exceeding

$$\bar{\kappa} = \left(\frac{4\pi^2 C^{5/2} K}{3} \right)^{1/2} N^{-1/4} \left[\iint_{S^2} B(x, y) dx dy \right]^{-1/4}. \quad (11)$$

It is easy to see that for sufficiently small ρ the weights $a_j = \langle g^j, g_j \rangle (\langle g_h, g_j \rangle)^{-1}$ are positive.

The estimate (11) is not overlapped by an estimate in the Monte Carlo method [9], where the error is estimated on a fixed function on $L^2(S^2)$. The realizations of the random field are formal series of spherical functions and are not bound to belong to $L^2(S^2)$ (see also paragraph 1).

11. Let

$$B(\theta) \in C^2[0, \pi], \quad K = \max_{\theta \in [0, \pi]} |B_{\theta\theta}''(\theta)|.$$

In this case $B_{\theta}'(0) = 0$ and the estimate of (11) can be improved by estimating the Taylor series for B :

$$|B(x, y) - B(x_j, x_j)| \leq K \frac{[\theta(x, x_j) + \theta(y, x_j)]^2}{2},$$

$$\left\{ \iint_{U_j, U_j} B(x, y) dx dy - \left[\int_{U_j} B(x_j, x) dx \right]^2 [B(x_j, x_j)]^{-1} \right\}^{1/2} \leq \frac{\pi\rho_j^3}{3} (13K)^{1/2}.$$

Instead of (10) and (11) the following estimates hold:

$$\min_{a_j \in \mathbb{C}} \left\| g_0 - \sum_{j=1}^N a_j g_j \right\| \leq \frac{N\pi\rho^3}{3} (13K)^{1/2}$$

and

$$\bar{x} = \left(\frac{13\pi^2 C^3 K}{9} \right)^{1/2} N^{-1/2} \left[\int_{S^1} \int_{S^1} B(x, y) dx dy \right]^{-1/2}.$$

The assertions of paragraphs 10, 11 remain true if the Lipschitz property or differentiability of the function B is assumed only on the segment $[0, \epsilon_0]$ and integrability on the segment $[0, \pi]$.

12. Let the functional (2) act not on the field φ itself, but on a homogeneous random field η , homogeneously connected with φ (noise is applied to the field φ). In this case it is easy to write down the analog of system (2).

13. Similar investigations can be carried out on the homogeneous space of any compact Lie group. The corresponding definitions and theorems on homogeneous random fields are given in [2, 3].

Translated by J. Berry.

REFERENCES

1. GANDIN, L. S. *Objective analysis of meteorological fields* (Ob'ektivnyi analiz meteorologicheskikh polei), Gidrometeoizdat, Leningrad, 1963.
2. YAGLOM, A. M. Positive-definite functions and homogeneous random fields on groups and homogeneous spaces. *Dokl. Akad. Nauk SSSR*, 135, 6, 1342-1346, 1960.
3. HANNAN, E. *Group representations and applied probability* (Predstavlenniya grupp i priklannaya teoriya veroyatnostei), "Mir", Moscow, 1970.
4. SOBOLEV, S. L. *Introduction to the theory of cubature formulas* (Vvedenie v teoriyu kubaturnykh formul), "Nauka", Moscow, 1974.
5. STROUD, A. N. *Approximate calculation of multiple integrals*. Prentice-Hall, New Jersey, 1971.
6. LEBEDEV, V. I. Values of the nodes and weights of ninth to seventeenth order Gauss-Markov quadrature formulae invariant under the octahedron group with inversion. *Zh. vychisl. Mat. mat. Fiz.*, 15, 1, 48-54, 1975.
7. PONTRYAGIN, L. and SHNIREL'MAN, L. On a metrical property of dimensionality. Appendix to the book: HUREWICZ, W. and WALLMAN, H. *Dimension theory* (Teoriya razmernosti), Izd-vo in. lit., Moscow, 1948.
8. TOTH, L. F. *Expansions on a plane, on a sphere and in space* (Raspolozheniya na ploskostim na sfere i v prostranstve), Fizmatgiz, Moscow, 1958.
9. SOBOL, I. M. *Numerical Monte Carlo methods* (Chislennyye metody Monte-Karlo), "Nauka", Moscow, 1973.

EFFICIENT MONTE CARLO ALGORITHMS FOR EVALUATING THE CORRELATION CHARACTERISTICS OF CONDITIONAL MATHEMATICAL EXPECTATIONS*

G. A. MIKHAILOV

Novosibirsk

(Received 10 March 1975)

IN THE proposed algorithms only two samples of the conditional distribution are used for each value of a condition. It is shown that in this way it is possible to calculate efficiently the correlation characteristics of the solution of the particle transfer equation with random coefficients and the optimal parameters for the "splitting method".

We consider the random vector quantity

$$\xi = \xi(\omega, \sigma) = \{\xi_1(\omega, \sigma), \dots, \xi_n(\omega, \sigma)\},$$

where ω is a random point of some abstract space (for example, the trajectory of a Markov chain); a joint probability distribution is specified for ω and the random vector $\sigma = \{\sigma_1, \dots, \sigma_m\}$. We have to estimate the correlation (otherwise — the "autocorrelation") matrix for the random vector of the conditional mathematical expectations

$$I = I(\sigma) = M_\omega(\xi | \sigma)$$

and the "mutual-correlation" moments for the vectors I and σ , that is, the quantities

$$\begin{aligned} K[I_k, I_j] &= M_\sigma[(I_k - I_k^*)(I_j - I_j^*)], \quad k, j = 1, 2, \dots, n, \\ K[I_k, \sigma_i] &= M_\sigma[(I_k - I_k^*)(\sigma_i - \sigma_i^*)], \quad k = 1, 2, \dots, n, \quad i = 1, 2, \dots, m. \end{aligned}$$

Here $\sigma_i^* = M\sigma_i$, $I_k^* = MI_k$, the subscript of the sign of the mathematical expectation defines the distribution to which it corresponds.

The numerical characteristics indicated are easily computed by the Monte Carlo method, if for each sample value of σ the vector $I(\sigma)$ is determined exactly. It will be shown below that this can be done by estimating $I(\sigma)$ at random with respect to one (for mutual-correlation moments) or with respect to two (for autocorrelation moments) values of ω . For mutual-correlation moments this algorithm follows directly from the relation

$$\begin{aligned} K[I_k, \sigma_i] &= M_\sigma[(I_k - I_k^*)(\sigma_i - \sigma_i^*)] = M_\sigma[(\sigma_i - \sigma_i^*)M_\omega(\xi_k - I_k^* | \sigma)] \\ &= M_{(\sigma, \omega)}[(\sigma_i - \sigma_i^*)(\xi_k - I_k^*)] = K[\sigma_i, \xi_k], \end{aligned}$$

if the latter correlation moment exists. Let $\omega^{(1)}, \omega^{(2)}$ be independent sample values of ω for a specified value of σ and $\xi_k^{(1)} = \xi_k(\omega^{(1)}, \sigma)$, $\xi_j^{(2)} = \xi_j(\omega^{(2)}, \sigma)$. Then

$$\begin{aligned}
 K[\xi_k^{(1)}, \xi_j^{(2)}] &= M_{(\omega^{(1)}, \omega^{(2)}, \sigma)}[(\xi_k^{(1)} - I_k^*)(\xi_j^{(2)} - I_j^*)] \\
 &= M_\sigma[M_{\omega^{(1)}}(\xi_k^{(1)} - I_k^* | \sigma) M_{\omega^{(2)}}(\xi_j^{(2)} - I_j^* | \sigma)] = K[I_k, I_j].
 \end{aligned}$$

The representations obtained

$$K[I_k, \sigma_i] = K[\xi_k, \sigma_i], \quad K[I_k, I_j] = K[\xi_k^{(1)}, \xi_j^{(2)}]$$

may be called "randomized". It should be noted that randomization is widely used to construct efficient algorithms of the Monte Carlo method (see, for example, [1], 80, 94, 110, 116).

By a similar method we can estimate the leading central moments of the vector $I(\sigma)$ of order N by using N sample values of ω . Below we explain another method of calculating the autocorrelation moments, which may give more accurate estimates.

We first consider the estimate of the variance DI_k . The equation

$$D\xi_k = DI_k + M_\sigma D(\xi_k | \sigma), \quad (1)$$

is well known. Let $\xi_k' = (\xi_k^{(1)} + \xi_k^{(2)})/2$. It is obvious that

$$D\xi_k' = DI_k + 0.5M_\sigma D(\xi_k | \sigma). \quad (2)$$

From (1) and (2) we deduce the formula

$$DI_k = 2D\xi_k' - D\xi_k.$$

It happens that relations (1) and (2) are valid for an arbitrary autocorrelation moment also, that is,

$$\begin{aligned}
 K[\xi_k, \xi_j] &= K[I_k, I_j] + M_\sigma K(\xi_k, \xi_j | \sigma), \\
 K[\xi_k', \xi_j'] &= K[I_k, I_j] + 0.5M_\sigma K(\xi_k, \xi_j | \sigma).
 \end{aligned}$$

The last equations follow directly from the expression for the correlation moment and the properties of mathematical expectations. Accordingly,

$$K[I_k, I_j] = 2K[\xi_k', \xi_j'] - K[\xi_k, \xi_j]. \quad (3)$$

Therefore, all the correlation moments required can be estimated, by simulating for each value of σ two values of ω and calculating the statistical estimates of the corresponding moments for the vectors ξ and ξ' . It is obvious that the moments of ξ can be estimated by averaging the estimates of the corresponding mathematical expectations, obtained for $\omega^{(1)}$ and $\omega^{(2)}$.

Here the estimates of the moments for ξ and ξ' will be strongly dependent and the accuracy of the calculations by formula (3) may be high. Two applications of the algorithms explained will be considered below.

1. Calculation of the correlation characteristics of the solution of the transfer equation with random coefficients. Here we will suppose that σ is the vector of the random coefficients

of the transfer equation (for example, the scattering or absorption coefficients), ω is a random trajectory of a particle, and $I_k(\sigma)$ is some functional of the solution of the transfer equation, for example the flow of particles at a given point. The relation $I_k(\sigma) = M_{\omega} \xi_k(\omega, \sigma)$ means that ξ_k is an unbiased estimate for the quantity I_k . The relations obtained above show that for the estimation of the mutual and auto-correlation moments of the vector $I(\sigma)$ is not at all necessary to solve the transfer equation exactly for each realization of the vector σ ; these estimates can be obtained by simulating for each realization of σ only two trajectories altogether; estimates of the leading moments can be obtained by simulating the corresponding number of trajectories.

An approximate estimate of the moments sought can also be constructed by the standard method of linearization on the basis of the calculation of the corresponding derivatives by the Monte Carlo method. Algorithms for estimating the derivatives of the particle flow (intensity) from the coefficients of the transfer equation are presented and validated in [1]. We also note that in [2] approximate algorithms for estimating the moments of the intensity are constructed on the basis of perturbation theory for a statistically homogeneous model of the medium.

There is a series of problems of transfer theory for whose complete solution it is necessary to calculate the statistical moments of the intensity. These are, for example, the important problems of the scattering of light in the atmosphere, whose optical characteristics are subjected to continual random variations.

2. Estimation of the optimal parameters by the splitting method. Suppose it is necessary to calculate the mathematical expectation of the function $\xi(\omega, \sigma)$, that is, the quantity

$$I^* = M_{(\omega, \sigma)} \xi(\omega, \sigma).$$

For each realization of σ it is possible to simulate n independent values of ω and use the expression

$$I^* = M \xi^{(n)} = M n^{-1} \sum_{h=1}^n \xi(\omega_h, \sigma).$$

The variance of the random quantity $\xi^{(n)}$ is expressed by the formula

$$D \xi^{(n)} = A_1 + A_2/n,$$

where $A_1 = D I(\sigma)$, $A_2 = M_{\sigma} D(\xi^{(1)} | \sigma)$. The time required to obtain one realization of $\xi^{(n)}$ on the computer is

$$t^{(n)} = t_1 + t_2 n,$$

where t_1 corresponds to the choice of σ , and t_2 to the choice of ω . The well known optimal value of the parameter

$$n \approx \left(\frac{t_1}{t_2} \frac{A_2}{A_1} \right)^{1/2}$$

yields a minimum value of the derivative $t^{(n)} D \xi^{(n)}$. The direct estimation of the quantities A_1 , A_2 , t_1 and t_2 is difficult. However, it is possible to obtain statistical estimates of the variance and times for two values of the parameter n_1 and n_2 and solve the corresponding systems of linear equations, that is, use the equations

$$A_1 = \frac{1}{n_2 - n_1} (n_2 D\xi^{(n_2)} - n_1 D\xi^{(n_1)}),$$

$$A_2 = \frac{n_1 n_2}{n_2 - n_1} (D\xi^{(n_1)} - D\xi^{(n_2)}),$$

$$t_1 = \frac{n_2 t^{(n_1)} - n_1 t^{(n_2)}}{n_2 - n_1}, \quad t_2 = \frac{t^{(n_2)} - t^{(n_1)}}{n_2 - n_1}.$$

Here it is useful to correlate the sample of values $\xi^{(n_1)}, \xi^{(n_2)}$. By a similar method we can calculate the parameters of a multiple splitting, for which

$$D\xi = A_0 + \frac{A_1}{n^{(1)}} + \frac{A_2}{n^{(1)}n^{(2)}} + \dots + \frac{A_k}{n^{(1)}n^{(2)} \dots n^{(k)}},$$

$$t = t_0 + n^{(1)}t_1 + n^{(1)}n^{(2)}t_2 + \dots + n^{(1)}n^{(2)} \dots n^{(k)}t_k.$$

Here (see [3]) A_i is the mean value of the conditional variance corresponding to the i -th splitting, and t_i is the mean time of the realization of one experiment in the limits from the i -th to the $(i+1)$ -th splitting. It is useful to construct the simulation in such a way that the chain of splittings is as far as possible homogeneous and the following equations are satisfied:

$$\frac{A_i}{A_{i+1}} = a, \quad \frac{t_i}{t_{i+1}} = b, \quad i=0, 1, \dots, k.$$

Here (see [3]) the optimal values of the $n^{(i)}$ are identical: $n^{(1)} = \dots = n^{(k)} = n$. After calculating $D\xi$ and t for two values $n = n_1, n_2$, we obtain for a and b :

$$\frac{D\xi^{(n_1)}}{D\xi^{(n_2)}} = \left(1 + \sum_{i=1}^k \frac{a^i}{n_1^i}\right) \left(1 + \sum_{i=1}^k \frac{a^i}{n_2^i}\right)^{-1},$$

$$\frac{t^{(n_1)}}{t^{(n_2)}} = \left(1 + \sum_{i=1}^k n_1^i b^i\right) \left(1 + \sum_{i=1}^k n_2^i b^i\right)^{-1},$$

which are easily solved on a computer, after which the optimal value of n is determined by the formula $n = (b/a)^{1/k}$.

Translated by J. Berry.

REFERENCES

1. MIKHAILOV, G. A. *Some topics in the theory of Monte Carlo methods* (Nekotorye voprosy teorii metodov Monte-Karlo), "Nauka", Novosibirsk, 1974.
2. BELOV, V. F., GLAZOV, G. N. and KREKOV, G. M. Algorithms for calculating the fluctuations of laser signals scattered by clouds. In: *Monte Carlo methods in computational mathematics and mathematical physics* (Metody Monte-Karlo v vychisl. matem. i. matem. fiz.), 246-253, VTs SO Akad. Nauk SSSR, Novosibirsk, 1974.
3. OGIBIN, V. N. On the application of "splitting" and "tape measure" in calculations of particle transfer by the Monte Carlo method. In: *The Monte Carlo method in the problem of radiative transfer* (Metod Monte-Karlo v probleme perenosy izlucheniya), 72-82, Atomizdat, Moscow, 1967.

A SEARCH SCHEME FOR APPROXIMATE SOLUTIONS OF THE CONVEX PROGRAMMING PROBLEM*

V. Yu. LEBEDEV

Moscow

(Received 29 May 1975; revised 29 September 1975)

THE POSSIBILITY of using the weighted functional method to solve convex programming problems is proved. For problems with a strictly concave target function a modification of the method is proposed which in simple cases permits the difference between the value of the target function in the approximate solution generated and the actual optimum to be estimated.

Many papers have been published devoted to methods of solving mathematical programming problems by using the technique of smooth penalty functions. In particular, much attention has been given to the method of the modified Lagrange function, as one of the most promising for solving convex problems [1, 2]. The absence of the need to increase the penalty parameter so as to improve the solution, and a number of other general features associate it with the method of weighted functional, described in [3, 4].

The idea of solving a conditional maximum problem by the successive unconditional minimization of a weighted functional with step-by-step improvement of the estimate of the target function optimum, converging up to the actual optimum, is apparently due to Morrison. In [5] he proposed a similar scheme for a problem with constraints of the equality type. It was generalized to the case of inequalities in [6]. A scheme based on the same idea, but considerably faster and differing in the formula for recalculating the estimates, was proposed in [3], where with certain assumptions its local convergence in the non-linear problem to the conditional extremum was proved. For the linear case the global convergence of this scheme after a finite number of steps and its stability to inaccuracies of the solution of the auxiliary problems were established in [4]. In the present paper the global convergence and stability of the scheme are proved for the convex case.

We consider a problem of the following form:

$$f(x) \rightarrow \max, \quad \varphi(x) \leq 0. \quad (1)$$

Here $x \in R^n$, $f(x)$ is a scalar concave function in R^n ; $\varphi(x)$ is a vector function each of whose m coordinates is a scalar function convex in R^n . It is assumed that $f(x)$, $\varphi(x)$ are continuously differentiable and that the set of possible points of problem (1) is bounded and not empty. Below the maximum value of the target function in problem (1) is denoted by $f(x^*)$.

We introduce the so-called weighted functional:

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 249-254, 1977.

$$\Psi(x, d) = \frac{1}{2} [d - f(x)]_+^2 + \frac{1}{2} \sum_{i=1}^m [\varphi_i(x)]_+^2.$$

Here d is a scalar quantity which henceforth we will call the estimate of the target function of problem (1), and by definition, $z_+ = \max\{0, z\}$. The function $\Psi(x, d)$ is convex and continuously differentiable with respect to its own variables. Because of the boundedness of the permissible set of problem (1), the aggregate of points of the form $\{x : \varphi_+^2(x) \leq \sigma\}$ will be bounded for any non-negative values of $\sigma < +\infty$ (see [7]). This implies that for any d the function $\Psi(x, d)$ attains a maximum with respect to x on the compact. This minimum will be positive if $d > f(x^*)$, and equal to zero otherwise.

The following scheme of search for an approximate solution of problem (1) is proposed. Let there be given an initial estimate d_1 , a point x_0 and two accuracy parameters $\varepsilon_1 > 0$, $\varepsilon_2 > 0$. A point x_1 is sought for which

$$\Psi(x_1, d_1) \leq \Psi(x_0, d_1) \quad (2)$$

and either

$$\|\Psi_x(x_1, d_0)\| \leq \varepsilon_1 (d_0 - f(x_1)), \quad \Psi(x_1, d_0) > \varepsilon_2, \quad (3)$$

or

$$\Psi(x_1, d_0) \leq \varepsilon_2. \quad (4)$$

If (4) is satisfied, the process ceases. Otherwise the quantity d_2 is calculated:

$$d_2 = d_1 - \frac{\Psi(x_1, d_1)}{d_1 - f(x_1)} < d_1, \quad (5)$$

and beginning with d_2 , x_1 the procedure is repeated. The method of choosing the points x_k within the limits of the constraints (3), (4) is not specified here. In any case, any convergent algorithm for minimizing the function $\Psi(x, d)$ with respect to x ensures the constructive possibility of finding them.

The following theorem holds.

Theorem 1

For any values of d_1 , x_0 , $\varepsilon_1 > 0$, $\varepsilon_2 > 0$ the process (2)–(5) ceases after a finite number of steps, where the points of its stopping, obtained for fixed x_0 , $d_1 \geq f(x^*)$ and different ε_1 , ε_2 will as $\varepsilon_1, \varepsilon_2 \rightarrow 0$ converge to a set of solutions of problem (1).

Proof. We prove the first part of the assertion of the theorem indirectly. We assume that d_1 , x_0 , $\varepsilon_1 > 0$, $\varepsilon_2 > 0$ exist such that as a result of the realization of the process (2)–(5) an infinite sequence $\{d_k\}$, $\{x_k\}$, will be obtained for which

$$\Psi(x_k, d_k) > \varepsilon_2.$$

Formula (5) for the quantities d_{n+1} can be rewritten as follows:

$$d_{k+1} = f(x_k) - \frac{\|\varphi_+(x_k)\|_E^2}{d_k - f(x_k)} = f(x_k) - (p_k, \varphi(x_k)),$$

where

$$p_k = \frac{\varphi_+^T(x_k)}{d_k - f(x_k)} \geq 0.$$

Here

$$\|\Psi_x'(x_k, d_k)\|_E = (d_k - f(x_k)) \left\| \frac{df(x_k)}{dx} - p_k \frac{d\varphi(x_k)}{dx} \right\|_E \leq e_1 (d_k - f(x_k)). \quad (6)$$

We introduce the vector

$$\Delta c_k = \frac{df(x_k)}{dx} - p_k \frac{d\varphi(x_k)}{dx}.$$

It is obvious from (6) that $\|\Delta c_k\|_E \leq e_1$.

We consider a linear programming problem of the form

$$\begin{aligned} \left(\frac{df(x_k)}{dx} - \Delta c_k, x \right) &= \left(p_k \frac{d\varphi(x_k)}{dx}, x \right) \rightarrow \max, \\ \frac{d\varphi(x_k)}{dx} (x - x_k) + \varphi(x_k) &\leq 0. \end{aligned} \quad (7)$$

The vector p_k is a permissible vector of the problem, dual to (7). Because of the convexity of the functions $\varphi_i(x)$, any of the permissible vectors of problem (1) will satisfy the constraints of problem (6), that is, the permissible set of problem (6) and its dual are not empty. This implies that problem (6) has a solution. We denote it by x' . By duality

$$\left(p_k, \frac{d\varphi(x_k)}{dx} x_k - \varphi(x_k) \right) \geq \left(\frac{df(x_k)}{dx} - \Delta c_k, x' \right).$$

This inequality can only be strengthened if on its right side x' is replaced by any solution x^* of problem (1). Therefore, we have

$$\left(\frac{df(x_k)}{dx} - \Delta c_k, x^* \right) \leq \left(p_k, \frac{d\varphi(x_k)}{dx} x_k - \varphi(x_k) \right). \quad (8)$$

Also, the concavity of the function $f(x)$ implies that

$$\left(\frac{df(x_k)}{dx}, x^* \right) \geq f(x^*) - f(x_k) + \left(\frac{df(x_k)}{dx}, x_k \right).$$

Taking into account (8), we obtain from this

$$\left(p_h, \frac{d\varphi(x_h)}{dx} x_h - \varphi(x_h) \right) \geq f(x^*) - f(x_h) + \left(\frac{df(x_h)}{dx}, x_h \right) - (\Delta c_h, x^*).$$

Consequently,

$$\begin{aligned} d_{h+1} = f(x_h) - (p_h, \varphi(x_h)) &\geq f(x^*) + \left(\frac{df(x_h)}{dx} - p_h \frac{d\varphi(x_h)}{dx}, x_h \right) \\ - (\Delta c_h, x^*) &= f(x^*) + (\Delta c_h, x_h - x^*). \end{aligned} \quad (9)$$

Since the points x_k and x^* belong to a bounded set $\{x: \|\varphi_+(x)\|_E^2 \leq \Psi(x_0, d_1)\}$, a number $M > 0$, exists, independent of the number k and such that $\|x_h - x^*\|_E \leq M$. Therefore,

$$d_{h+1} \geq f(x^*) - \varepsilon_1 M.$$

Therefore the sequence $\{d_h\}$ decreases monotonically and is underbounded. Therefore,

$$\lim_{h \rightarrow \infty} (d_h - d_{h+1}) = 0,$$

and since $d_h - d_{h+1} \geq [\Psi(x_h, d_h)]^{1/2} \geq 0$, it is obvious that

$$\lim_{h \rightarrow \infty} \Psi(x_h, d_h) = 0.$$

This contradicts the initial supposition that $\Psi(x_h, d_h) > \varepsilon_2 > 0$. The contradiction proves the first assertion of the theorem. The second assertion is proved in exactly the same way as in the linear case considered in [4].

Theorem 1 shows that, possessing an increased estimate of the optimum of problem (1) and having chosen sufficiently small values $\varepsilon_1 > 0$, $\varepsilon_2 > 0$, it is possible to rely on being able to obtain a better approximate solution by the scheme of (2)–(5). If there is no satisfactory prior estimate, it is possible to begin the search for a solution by maximization of the ordinary penalty function

$$\Psi(x) = f(x) - \frac{1}{2} \varphi_+^2(x),$$

continuing the calculations until at the current point x_0 the condition $\|\Psi_x'(x_0)\|_E \leq \varepsilon_1$ is satisfied, after which, putting $d_1 = f(x_0) - \varphi_+^2(x_0)$, we return to the scheme (2)–(5). The "extended" scheme thus obtained, beginning with an arbitrary point, will cease for any $\varepsilon_1 > 0$, $\varepsilon_2 > 0$ after a finite number of steps, giving excellent solutions for small values of ε_1 , ε_2 .

It is obvious from the proof of Theorem 1 what will happen if in (3), (4) we put $\varepsilon_1 = \varepsilon_2 = 0$, that is, solve the intermediate problems of minimizing the weighted functional completely. Then beginning with $d_1 \geq f(x^*)$ the scheme of (2)–(5) either finds the exact solution of problem (1) after a finite number of steps (as, for example, in the linear case [4]), or constructs a maximizing sequence $\{x_h\}$ and a sequence $\{d_h\}$, converging to the optimum of problem (1) on the right. In the case where the function $f(x)$ is strictly concave, the inequalities $d_h \geq f(x^*)$ can be ensured even without solving the intermediate problems exactly. For this in the scheme (2)–(5) it is necessary to replace condition (3) of the recalculation of the estimate d_k by another of the following form:

$$\begin{aligned}
d_k &> f(x_k), \\
\|\Psi'_x(x_k, d_k)\|_E &\leq \varepsilon_1 \min \{ \|\varphi_+(x_k)\|_{E^2}, [\|\varphi_+(x_k)\|_{E^2} (d_k - f(x_k))]^{1/2} \}, \\
\Psi(x_k, d_k) &> \varepsilon_2 > 0.
\end{aligned} \tag{10}$$

The following theorem holds.

Theorem 2

Let the function $f(x)$ be twice continuously differentiable, let there exist a positive constant μ such that for any $x \in R^n$, $y \in R^n$ the inequality

$$\left(y^T \frac{d^2 f(x)}{dx^2}, y \right) \leq -2\mu \|y\|_{E^2},$$

is satisfied, and at x^* — the solution of problem (1) — let the derivative $df(x^*)/dx$ be non-zero. (For $df(x^*)/dx=0$ and small values of $\varepsilon \geq 0$ the system (10) is inconsistent.) Then for sufficiently small values of $\varepsilon_1 \geq 0$ and any $x_k, d_k, \varepsilon_2 > 0$, satisfying conditions (10), the value of d_{k+1} calculated by formula (5) will be not less than $f(x^*)$.

Proof. We take any $\varepsilon_1 \geq 0$, provided that

$$\varepsilon_1 = \min \left\{ \frac{\mu}{2\|df(x^*)/dx\|_E}, \left(\frac{\mu}{2} \right)^{1/2} \right\}, \tag{11}$$

is satisfied, and let $x_k, d_k, \varepsilon_2 > 0$ satisfy the inequalities (10). Then for d_{k+1} calculated by formula (5), it is possible taking into account the strict concavity of $f(x)$, to write down an analog of the estimate (9) obtained in the proof of Theorem 1:

$$d_{k+1} \geq f(x^*) + \mu \|x_k - x^*\|_{E^2} + (\Delta c_k, x_k - x^*).$$

Here

$$\|\Delta c_k\|_E \leq \varepsilon_1 \min \{ \delta_k, (\delta_k)^{1/2} \}, \quad \delta_k = \|\varphi_+(x_k)\|_{E^2} / [d_k - f(x_k)].$$

From this, allowing for (11), we obtain

$$\begin{aligned}
d_{k+1} &\geq f(x^*) + \mu \|x_k - x^*\|_{E^2} - \varepsilon_1 \min \{ \delta_k, (\delta_k)^{1/2} \} \|x_k - x^*\|_E \\
&\geq f(x^*) + \mu \|x_k - x^*\|_{E^2} - \min \left\{ \frac{\mu \delta_k}{2\|df(x^*)/dx\|_E}, \left(\frac{\mu}{2} \delta_k \right)^{1/2} \right\} \|x_k - x^*\|_E.
\end{aligned}$$

Moreover, by the concavity of $f(x)$,

$$\begin{aligned}
\|df(x^*)/dx\|_E \|x_k - x^*\|_E &\geq (df(x^*)/dx, x_k - x^*) \\
&\geq f(x_k) - f(x^*) = f(x_k) - d_{k+1} + d_{k+1} - f(x^*) \\
&\geq \delta_k - \min \left\{ \frac{\mu \delta_k}{2\|df(x^*)/dx\|_E}, \left(\frac{\mu}{2} \delta_k \right)^{1/2} \right\} \|x_k - x^*\|_E.
\end{aligned}$$

Consequently,

$$\begin{aligned}
& \mu \|x_k - x^*\|_E^2 \\
& \geq \delta_k \mu \|x_k - x^*\|_E \left(\left\| \frac{df(x^*)}{dx} \right\|_E + \min \left\{ \frac{\mu \delta_k}{2 \|df(x^*)/dx\|_E}, \left(\frac{\mu}{2} \delta_k \right)^{1/2} \right\} \right)^{-1} \\
& \geq \min \left\{ \frac{\mu \delta_k}{2 \|df(x^*)/dx\|_E}, \left(\frac{\mu}{2} \delta_k \right)^{1/2} \right\} \|x_k - x^*\|_E,
\end{aligned}$$

that is,

$$d_{k+1} \geq f(x^*), \quad (12)$$

which is what it was required to prove.

The scheme (2), (10), (4), (5) will be finite for any $d_1, x_0, \varepsilon_2 > 0$ and any sufficiently small ε_1 (by (12)), provided that the target function of problem (1) is strictly concave and $df(x^*)/dx \neq 0$. Then the intermediate problems can be solved by any convergent algorithm for the minimization of $\Psi(x, d)$ with respect to x . Therefore, if we are able efficiently to estimate the quantities $\mu, \|df(x^*)/dx\|$, downwards and upwards, respectively, (as, for example, in the problem of finding the minimum of the distance from a given point to a polyhedron), then, having $d_1 \geq f(x^*)$ and using the scheme (2), (10), (4), (5), it is possible to find an approximate solution for which the norm of the discrepancies of the constraints, and also the difference between the actual optimum and the value of the target function obtained do not exceed the square root of the chosen value of the parameter ε_2 .

Translated by J. Berry.

REFERENCES

1. POLYAK, B. T. and TRET'YAKOV, N. V. An iterative method of linear programming. *Ekonomika matem. metody*, 8, 5, 740-751, 1972.
2. TRET'YAKOV, N. V. The method of penalty estimates for convex programming problems. *Ekonomika matem. metody*, 9, 3, 526-540, 1973.
3. KOWALIK, J., OSBORNE, M. R. and RYAN, D. M. A new method for constrained optimization problems. *Operat. Res.*, 17, 6, 973-983, 1969.
4. LEBEDEV, V. Yu. An approximate algorithm for the solution of the linear programming problem. *Zh. vychisl. Mat. mat. Fiz.*, 14, 4, 1052-1058, 1974.
5. MORRISON, D. Optimization by least squares. *SIAM J. Numer. Analysis*, 5, 1, 83-88, 1968.
6. BELYAEVA, A. R., OSTROVSKII, G. M. and BEREZHINSKII, T. A. A method of solving a problem at a conditional extremum. *Avtomatika vychisl. tekhn.*, No. 4, 56-65, 1974.
7. FIACCO, A. V. and McCORMICK, G. P. *Non-linear programming: Sequential unconstrained minimization techniques* (Nelineinoe programmirovaniye. Metody posledovatel'noi bezuslovnoi minimizatsii), "Mir", Moscow, 1972.

DUALITY IN MULTI-TARGET PROGRAMMING*

V. D. NOGIN

Leningrad

(Received 14 March 1975)

A DUAL problem of finite-dimensional multi-target programming is constructed for the most general assumptions on the vector functions to be maximized and the constraints. The concave case is considered in detail.

Extremal problems of multi-target programming, in which the maximization is performed not with respect to one, but with respect to several target functions, have recently been intensively studied. A whole set of maximal (efficient, Pareto optimal, non-improvable) elements usually emerges as the solution of such problems. At the present time the theory of multi-target programming has not yet arrived at the same stage of development as the ordinary theory of mathematical programming, however it appears to contain fairly interesting and difficult mathematical problems [1-4] and is of interest from the point of view of applications in the most diverse fields: in economics, the theory of games, the theory of optimal decision making and in all problems of the choice of optimal solutions with incongruent criteria.

It is known that one of the most important ideas of mathematical programming theory is the idea of duality, by which a correspondence is established between the original extremal problem and the dual problem closely associated with it. A joint study of both problems is fruitful both for the development of numerical algorithms and also for qualitative investigations of extremal problems. At the present time a comparatively detailed study has been made of duality in ordinary mathematical programming (see, for example, [5]). In multi-target programming duality was considered only in [6, 7]. The dual problem of finding maximal elements in the completely linear case was first formulated in [6]. In [7] a dual problem was constructed from assumptions about concavity and differentiability of the target vector functions and constraints.

In the present paper the dual problem is formulated for the most general assumptions by means of a vector Lagrange function and a certain non-transitive binary relation on a set of values of the Lagrange vector function. The present treatment of duality differs from the approach in [7] and includes as a particular case (when the target vector function degenerates into a scalar function) the approach developed in [8]. In the case where the vector function to be maximized and the constraints are concave (and not necessarily differentiable), conditions are given for which the set of solutions of the direct problem is identical with the set of solutions of the dual problem. It is shown that in the linear case the dual problem is similar to the dual problem considered in [6].

1. Let $a = (a_1, \dots, a_m)$, $b = (b_1, \dots, b_m)$. We agree that

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 254-258, 1977.

$$a \geq b \Leftrightarrow a_i \geq b_i, \quad i=1, 2, \dots, m,$$

$$a \geq b \Leftrightarrow a \geq b \text{ or } a \neq b,$$

$$a > b \Leftrightarrow a_i > b_i, \quad i=1, 2, \dots, m,$$

$$a \leq b \Leftrightarrow b \geq a.$$

It is obvious that the relation $a \leq b$ is satisfied if and only if either $a = b$, or a subscript $i \in \{1, 2, \dots, m\}$, is found such that $a_i > b_i$. It should be noted that the binary relation \leq is not transitive for $m > 1$.

Definition 1. The finite element $a^0 \in A$, $A \subset E^m$, is called the maximal element of the set A , if the satisfaction for some $a \in A$ of the inequality $a \geq a^0$ implies that $a = a^0$.

Definition 2. The element $a^0 \in A$ is called minimal, if $-a^0$ is the maximal element of the set $-A$.

It is easy to verify that the following definition is equivalent to definition 1.

Definition 1'. A finite element $a^0 \in A$ is the maximal element of the set A if $a^0 \leq a$ is satisfied for any $a \in A$.

We will suppose that the vector functions $F(x) = (f_1(x), \dots, f_m(x))$, $G(x) = (g_1(x), \dots, g_k(x))$ are defined on the set $X \subset E^n$. We put

$$D = \{x \in X | G(x) \geq 0_k\},$$

$$L(x, \lambda) = (L_1(x, \lambda), \dots, L_m(x, \lambda)),$$

$$L_j(x, \lambda) = f_j(x) + (\lambda, G(x)), \quad j=1, 2, \dots, m,$$

$$\Lambda = \{\lambda \in E^k | \lambda \geq 0_k\},$$

where $(\lambda, G(x))$ denotes the scalar product of the vector λ and $G(x)$. We will also consider that the vector Lagrange function $L(x, \lambda)$ is defined on the set $X \times \Lambda$ and $D \neq \emptyset$.

2. **Direct problem I.** Find the maximal elements of the set

$$P = \bigcup_{x \in X} P(x), \quad (1)$$

where

$$P(x) = \bigcap_{\lambda \in \Lambda} \{p \in E^m | p \leq L(x, \lambda)\}.$$

The solutions of this problem are identical with the maximal elements of the set

$$\left\{ \bigcup_{x \in D} P(x) \right\} \cup \left\{ \bigcup_{x \in X \setminus D} P(x) \right\}.$$

For $x \in X \setminus D$ every component of the vector $p \in P(x)$ is not underbounded, therefore the set of solutions of the direct problem is identical with the set of maximal elements of the set $\bigcup_{x \in D} P(x)$. Also, as is easily verified, if $p^0 \in P(x^0)$ is a solution of the direct problem,

$x \in D$

then $p^0 = F(x^0)$ for some $x^0 \in D$. Therefore the direct problem I is equivalent to the direct problem II.

Direct problem II. Find the set of maximal elements of the set

$$\bigcup_{x \in D} F(x). \quad (2)$$

In this form the direct problem has the standard form of the multi-target programming problem of the discovery of efficient elements.

Dual problem I. Find the minimal elements of the set

$$H = \bigcup_{\lambda \in \Lambda} H(\lambda), \quad (3)$$

where

$$H(\lambda) = \bigcap_{x \in X} \{h \in E^m \mid h \preceq L(x, \lambda)\}.$$

Lemma 1

The relation $a \preceq b$ for $a, b \in E^m$ is satisfied if and only if a vector $\mu > 0_m$, exists for which $(\mu, a) \geq (\mu, b)$ and

$$\sum_{i=1}^m \mu_i = 1. \quad (4)$$

Proof. Necessity. Let

$$E^- = \{x \in E^m \mid x \leq 0_m\}.$$

By the condition of the lemma the point $a-b$ does not belong to the convex set E^- . It is easy to understand that the set E^- may be separated from the point $a-b$ by some hyperplane $(\mu, x) = 0$, where

$$(\mu, x) < 0 \quad \forall x \in E^-, \quad (\mu, a-b) \geq 0.$$

The first inequality and the definition of the set E^- imply $\mu > 0_m$. It is obvious that the vector μ can be so chosen that the second inequality in (4) is satisfied.

Sufficiency. If $(\mu, a) \geq (\mu, b)$, then by the inequality $\mu > 0_m$, the assumption $a \leq b$ implies the contradiction: $(\mu, a) < (\mu, b)$.

By the lemma proved the dual problem I is equivalent to the dual problem II.

Dual problem II. Find the minimal elements of the set (3), where

$$H(\lambda) = \bigcap_{x \in X} \bigcup_{\mu \in M} \{h \in E^m \mid (\mu, h) \geq (\mu, F(x)) + (\lambda, G(x))\},$$

$$M = \left\{ \mu \in E^m \mid \mu > 0_m, \sum_{i=1}^m \mu_i = 1 \right\}.$$

We will now explain the connection between the direct and dual problems.

Lemma 2

For any $p \in P, h \in H$ the relation $p \geq h$ holds.

Proof. We assume the contrary: for some pair of elements $p^0 \in P, h^0 \in H$ the inequality $p^0 \geq h^0$ holds. Let $p^0 \in P(x^0), h^0 \in H(\lambda^0)$, where $x^0 \in X, \lambda^0 \in \Lambda$. By the definition of the sets $P(x), H(\lambda)$ we obtain

$$L(x^0, \lambda^0) \geq p^0 \geq h^0 \leq L(x^0, \lambda^0),$$

which implies the contradiction: for some $i \in \{1, 2, \dots, m\}$ we obtain $L_i(x^0, \lambda^0) > L_i(x^0, \lambda^0)$.

The relation $p \geq h$ holds both for the solutions p^0 of the direct and for the solutions h^0 of the dual problems, that is, $p^0 \geq h^0$, which is the vector analog of the known inequality $\max \min L(x, \lambda) \leq \min \max L(x, \lambda)$ for the scalar Lagrange function.

Corollary. The set of elements, which are simultaneously solutions of the direct and dual problems, is of the form $P \cap H$. The necessity of the assertion is obvious, and the sufficiency follows from Lemma 2 and the definition 1' of the maximal (minimal) element.

3. Here continuity and concavity of the vector functions $F(x), G(x)$ on the convex set X will be assumed.

Definition 3. The maximal element a^0 of the set A is said to be properly maximal, if a vector $\mu \in M$ exists such that the inequality $(\mu, a^0) \geq (\mu, a)$ is satisfied for all $a \in A$.

Remark 1. Usually by a properly maximal element is understood the value of the vector function F at a properly efficient point [2]. But, as shown in [2], when the assumptions of continuity and concavity are satisfied this concept is equivalent to the properly maximal element of definition 3.

Lemma 3

We suppose that a point $x \in X$ has been found such that $g_j(x) > 0$ for all non-linear functions g_j . Then every properly maximal element of the set P is a solution of the dual problem.

Proof. Let p^0 be a properly maximal element of the set P . Because of the maximality, $p^0 = F(x^0)$ for some $x^0 \in D$, and since p^0 is properly maximal, then by the results of [2], vectors $\mu^0 \in M, \lambda^0 \in \Lambda$ can be found such that for all $x \in X$ the inequality

$$(\mu^0, F(x^0)) \geq (\mu^0, L(x, \lambda^0)).$$

is satisfied. Accordingly, $p^0 \in P \cap H$, and applying the corollary of Lemma 2, we obtain that p^0 is a solution of the dual problem.

Theorem 1

Let the regularity condition of Lemma 3 be satisfied. We will assume: a) every maximal element of the set P is properly maximal or b) the set X is closed, there exists at least one

properly maximal element of the set P and the limit of any convergent sequence of solutions of the dual problem belongs to the set H . Then every solution of the direct problem is a solution of the dual problem.

Proof. In the case when condition a) is satisfied Lemma 3 works. Let b) hold. It is obvious that the set D is convex and closed, thereby by the remark of [2], p. 623–624, for a concave and continuous vector function $F(x)$ the set of maximal elements of P occurs in the closure of the set of properly maximal elements of P . The limit of any convergent sequence of properly maximal elements belongs to the closed set P , and thanks to Lemma 3 and the hypothesis of the theorem, belongs to the set H . Consequently, every solution of the direct problem is a solution of the dual problem.

Remark 2. It is obvious that when $m = 1$ condition a) is necessarily satisfied. V. V. Podinovskii has proved that for $m \geq 1$ condition a) holds for linear F, G (this result was also published in [9]). The author has established (because of the unwieldiness of the proof it is impossible to include it here) the proper maximality of the maximal elements in the non-linear case, when F is simultaneously concave and pseudo-concave, and G is simultaneously quasi-concave and pseudo-convex.

Theorem 2

Let the regularity condition hold, the set X be closed and let there exist at least one properly maximal element of the set P . Then every solution of the dual problem is a solution of the direct problem.

Proof. We suppose that $h^0 \in H$ is a solution of the dual problem. Since h^0 is a minimal element of the set H , we have $h^0 \in \text{int } H$.

We prove the inclusion $\bar{P} \subset \text{int } H$, where \bar{P} denotes the complement of P to E^m . Because of the assumptions of the theorem the set P is convex and closed. Hence, if $h' \in \bar{P}$, then as $\varepsilon > 0$ exists such that the sphere $B(h')$ of radius with centre at the point h' does not intersect the set P . In this case the sets P and $B(h')$ may be strongly separate, that is, there exists a non-zero vector $\mu \in E^m$ such that

$$(\mu, h) > \alpha \quad \forall h \in B_\varepsilon(h'), \quad (\mu, p) < \alpha \quad \forall p \in P.$$

By the definition of the set P we obtain from the second inequality $\mu \geq 0_m$. We may obviously consider that (4) is satisfied. Let p^0 be a properly maximal element of the set P and $(\mu^0, p^0) \geq (\mu^0, p)$ for all $p \in P$ and for some $\mu^0 \in M$. We consider the vector

$$\mu^\omega = \omega \mu^0 + (1 - \omega) \mu,$$

belonging to the set M for any $\omega \in (0, 1)$. Obviously,

$$\begin{aligned} (\mu^\omega, p^0) &= \omega (\mu^0, p^0) + (1 - \omega) (\mu, p^0), \\ (\mu^\omega, h) &= \omega (\mu^0, h) + (1 - \omega) (\mu, h). \end{aligned}$$

As a consequence of the fact that the linear function (μ^0, h) is bounded on the set $B(h')$ and $(\mu, h) > (\mu, p^0)$ for all $h \in B_\varepsilon(h')$, the positive ω can be taken so small that for all $h \in B_\varepsilon(h')$ the inequality

$$(\mu^\omega, h) \geq (\mu^\omega, p^0).$$

is satisfied. By Lemma 3, p^0 is a solution of the dual problem, and hence $p^0 \in H$. This together with the above inequality gives $h' \in \text{int } H$.

By the inclusion proved, $h^0 \in P \cap H$, which together with the corollary to Lemma 2 indicates the membership of the element h^0 is the set of solutions of the direct problem.

We consider the linear case. Let $X = E^n$, $F(x) = Cx$, $G(x) = b - Ax$, where C and A are $m \times n$ and $k \times n$ matrices respectively. Using the fact of the coincidence in the linear case of the maximal and properly maximal elements, and of the necessary and sufficient conditions of proper maximality [2], we arrive at the following formulation of the dual problem: find the minimal elements of the set

$$\bigcup_{\lambda \in \Lambda} \bigcup_{\mu \in M} \{h \in E^m \mid (\mu, h) \cong (\lambda, b)\}$$

subject to the condition $\mu C = \lambda A$. In this form the linear dual problem was formulated in [6].

Translated by J. Berry;

REFERENCES

1. HURWICZ, L. Programming in linear topological spaces. In: ARROW, K. J., HURWICZ, L. and UDZAWA, H. *Studies in linear and non-linear programming* (Issledovaniya po lineinomu i nelineinomu programmirovaniyu), 65-155, Izd-vo in. lit., Moscow, 1962.
2. GEOFFRION, A. M. Proper efficiency and the theory of vector maximization. *J. Math. Analysis Applic.*, **22**, 3, 618-630, 1968.
3. Da CUNHA, N. O. and POLAK, E. Constrained minimization under vector-valued criteria in finite dimensional spaces. *J. Math. Analysis Applic.*, **19**, 1, 103-124, 1967.
4. Da CUNHA, N. O. and POLAK, E. Constrained minimization under vector-valued criteria in topological spaces. In: *Math. Theory of Control*, 96-108, Acad. Press, New York-London, 1967.
5. GOL'SHTEIN, E. G. *Theory of duality of mathematical programming and its applications* (Teoriya dvoistvennosti matematicheskogo programmirovaniya i ee prilozheniya), "Nauka", Moscow, 1971.
6. GALE, D., KUHN, H. W. and TUCKER, A. W. Linear programming and the theory of games. In: *Activity Anal. Production and Allocation*, 317-329, Wiley, New York, 1951.
7. SHONFELD, P. Some duality theorems for the non-linear vector maximum problem. *Unternehmensforschung*, **14**, 1, 51-63, 1970.
8. ZANGWILL, W. I. *Nonlinear programming: A unified approach* (Nelineinoe programmirovaniye. Edinyi podkhod), "Sov. radio", Moscow, 1974.
9. ISERMANN, H. Proper efficiency and the linear vector maximum problem. *Operat. Res.*, **22**, 1, 189-191, 1974.

ALGORITHM FOR SOLVING THE LINEAR PROGRAMMING PROBLEM BY THE LOADED FUNCTIONAL METHOD*

A. P. ABRAMOV and Yu. P. IVANILOV

(Received 25 March 1976)

A MODIFICATION of the loaded functional method proposed in [1, 2], using the trajectory of the local minima, is discussed. The algorithm finds the solution of the linear programming problem after a finite number of steps.

Consider the linear programming problem in the form

$$cx \rightarrow \max, \quad Ax \leq b. \quad (1)$$

Here $c^T, x \in R^n, b \in R^m$, and A is an $m \times n$ matrix. It is assumed that the vector c is non-zero and the set of solutions of problem (1) is not empty. The problem dual to (1) has the form

$$pb \rightarrow \min, \quad pA = c, \quad p \geq 0, \quad \text{where } p^T \in R^m. \quad (2)$$

We take the arbitrary number ω and construct the loaded functional (see [1, 2]):

$$\psi(x, \omega) = (\omega - cx)_+^2 + \sum_{i=1}^m (a_i x - b_i)_+^2, \quad (3)$$

where $z_+ = \max\{0, z\}$ is a truncation of the function z , and a_i is the i -th row of the matrix A . For a fixed value of ω the function $\psi(x, \omega)$ is convex and continuously differentiable in R^n , unbounded by zero. For any ω there exists

$$f(\omega) = \min_{x \in R^n} \psi(x, \omega). \quad (4)$$

This minimum equals zero if ω does not exceed the optimal value $\omega^* = cx^*$ of the functional of problem (1). Any point x at which a minimum is attained is a permissible solution of problem (1). But if $\omega > cx^*$, then $f(\omega)$ is a positive and monotonically increasing function. Moreover, because of the convexity of $\psi(x, \omega)$ with respect to the ensemble of variables, the function $f(\omega)$ is convex. For $\omega > cx^*$ at any point x where loaded functional (3) attains a minimum the inequality

$$\omega - cx > 0. \quad (5)$$

is satisfied.

Therefore, problem (1) reduces to a search for the largest root ω^* of the function $f(\omega)$. This root could be found by Newton's method. However, the convergence of this method in the

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 259-263, 1977.

neighborhood of ω^* is small, since $f(\omega)$ vanishes at $\omega = \omega^*$ together with its derivative. The method proposed in [3] gives at each iteration an improved value of the root of the function $f(\omega)$ with a step twice as great as in Newton's method, and in the neighborhood of ω^* this method is identical with the second-order Newton method. In both Newton's method and the method proposed in [3] it is necessary at the k -th step of the algorithm to determine the value $f(\omega^k)$ of minimization of the function $\psi(x, \omega^k)$, which for a large dimension of problem (1) is an extremely laborious operation. Below we present an algorithm for finding the largest root of the function $f(\omega)$ using the trajectory of the local minima of $x(\omega)$, where by definition

$$\psi(x(\omega), \omega) = \min_{x \in R^n} \psi(x, \omega), \quad (6)$$

which at each step permits us to solve a problem of smaller dimension than problem (4).

Let some number $\tilde{\omega} > c x^*$ and the vector $\tilde{x} = x(\tilde{\omega})$ be given. We assume that the point \tilde{x} is such that in some neighborhood of it the set of constraints $I \in \{1, 2, \dots, m\}$ of problem (1), and only this one, is violated, that is,

$$a_i \tilde{x} - b_i > 0, \quad i \in I, \quad a_i \tilde{x} - b_i < 0, \quad i \notin I. \quad (7)$$

Then in this neighborhood the point \tilde{x} of the loaded functional (3) taking into account (5) and (7) is written as follows:

$$\psi(x, \tilde{\omega}) = (\tilde{\omega} - cx)^2 + \sum_{i \in I} (a_i x - b_i)^2.$$

The necessary and sufficient condition for a minimum of the function $\psi(x, \tilde{\omega})$ at the point \tilde{x} is the vanishing of the derivative $\psi_x'(\tilde{x}, \tilde{\omega})$ at this point. We represent this condition in the form

$$\left(c^T c + \sum_{i \in I} a_i^T a_i \right) \tilde{x} = c^T \tilde{\omega} + \sum_{i \in I} b_i a_i^T. \quad (8)$$

Condition (8) specifies explicitly the trajectory $x(\omega)$, which is called the trajectory of local minima [4]. Therefore, instead of solving problem (4) we can solve the system of equations

$$(\omega - cx)_+ + c + \sum_{i=1}^m (a_i x - b_i)_+ + a_i = 0. \quad (9)$$

We calculate the value of the function $\psi(x, \omega)$ at the point $(\tilde{x} + \rho \xi, \tilde{\omega} + \rho)$, where the n -dimensional vector ξ defines the direction of the shift in the space of the x -es. By choosing the negative number ρ sufficiently small in modulus at the point $\tilde{x} + \rho \xi$ one and only one set of constraints \tilde{I} will be violated. Performing elementary transformations, we obtain

$$\begin{aligned} \psi(\tilde{x} + \rho \xi, \tilde{\omega} + \rho) - \psi(\tilde{x}, \tilde{\omega}) &= 2 \left[-(\tilde{\omega} - c\tilde{x})c + \sum_{i \in \tilde{I}} (a_i \tilde{x} - b_i) a_i \right] \rho \xi \\ &+ 2(\tilde{\omega} - c\tilde{x})\rho + \left[\sum_{i \in \tilde{I}} (a_i \xi)^2 + (1 - c\xi)^2 \right] \rho^2, \end{aligned}$$

from which by condition (9) we have

$$\psi(\bar{x} + \rho \tilde{\xi}, \tilde{\omega} + \rho) - \psi(\bar{x}, \tilde{\omega}) = 2(\tilde{\omega} - c\bar{x})\rho + \left[\sum_{i \in \bar{I}} (a_i \tilde{\xi})^2 + (1 - c\tilde{\xi})^2 \right] \rho^2. \quad (10)$$

It follows from (10) that for negative values of ρ sufficiently small in modulus, the expression on the right in (10) is negative. Therefore, the problem arises of choosing the best direction $\xi \in R^n$ such that

$$\sum_{i \in \bar{I}} (a_i \xi)^2 + (1 - c\xi)^2 \rightarrow \min. \quad (11)$$

This problem is equivalent to solving the system of equations

$$\left(c^T c + \sum_{i \in \bar{I}} a_i^T a_i \right) \xi = c^T. \quad (12)$$

We differentiate the system of equations (8) with respect to ω :

$$\left(c^T c + \sum_{i \in \bar{I}} a_i^T a_i \right) \frac{dx}{d\omega} = c^T. \quad (13)$$

Comparison of system (12) and (13) shows that the optimal value $\xi = \tilde{\xi}$ is identical with $dx/d\omega$, the solution of system (13).

Remark 1. We note that the derivative $dx/d\omega$ does not always exist. For some values of ω the left and right derivatives have different values (see below).

Taking into account (12), relation (10) reduces to the form

$$\psi(\bar{x} + \rho \tilde{\xi}, \tilde{\omega} + \rho) - \psi(\bar{x}, \tilde{\omega}) = 2(\tilde{\omega} - c\bar{x})\rho + (1 - c\tilde{\xi})\rho^2. \quad (14)$$

We note that (10) and (14) imply that $1 - c\tilde{\xi} \geq 0$. Multiplying the transposed system (12) for $\xi = \tilde{\xi}$ by $\tilde{\xi}$ and transforming, we have

$$\sum_{i \in \bar{I}} (a_i \tilde{\xi})^2 = (1 - c\tilde{\xi}) c\tilde{\xi}.$$

Consequently, $0 \leq c\tilde{\xi} \leq 1$ and $0 \leq 1 - c\tilde{\xi} \leq 1$. It is easy to show that

$$\begin{aligned} [\tilde{\omega} + \rho - c(\bar{x} + \rho \tilde{\xi})]^2 - (\tilde{\omega} - c\bar{x})^2 &= (1 - c\tilde{\xi}) \Delta\psi, \\ \sum_{i \in \bar{I}} [a_i(\bar{x} + \rho \tilde{\xi}) - b_i]^2 - \sum_{i \in \bar{I}} (a_i \bar{x} - b_i)^2 &= c\tilde{\xi} \Delta\psi, \end{aligned}$$

where $\Delta\psi$ denotes the left side in (14), that is, the discrepancies of each of the two terms of the functional $\psi(x, \omega)$ are reduced for $\rho < 0$.

We specify $\rho = \omega - \tilde{\omega}$ in such a way that the point $x = \tilde{x} + \rho \tilde{\xi}$ violates the set of constraints \tilde{I} , and it alone. Substitution of ω and x into system (8) shows that the point x lies on the trajectory of local minima of the function $\psi(x, \omega)$. Accordingly, this trajectory is linear, the kinks corresponding to the vanishing of $a_i x - b_i$.

This algorithm realizes the approximation to the optimal point x^* along the trajectory of local minima. We will give a formal description of it.

Step 0. For the specified $\omega^0 > c x^*$ calculate the initial point x^0 :

$$\psi(x^0, \omega^0) = \min_{x \in R^n} \psi(x, \omega^0).$$

Step 1. Put $k = 0$.

Step 2. Determine the set of violated constraints I^k of problem (1) at the point x^k .

Step 3. At the point ω^k calculate the left derivative

$$(dx/d\omega) |_{\omega^k - 0} = \xi^k.$$

It can be determined by solving system (13) or the equivalent problem (11), which is of smaller dimension, than problem (4) for finding the minimum of the function $\psi(x, \omega)$.

Step 4. Put

$$x^{k+1} = x^k + (\omega^{k+1} - \omega^k) \xi^k,$$

where

$$\omega^{k+1} = \max_{i \in I^k} \omega_i^{k+1},$$

and ω_i^{k+1} are determined from the relations

$$a_i [x^{k+1} + (\omega_i^{k+1} - \omega^k) \xi^k] - b_i = 0, \quad i \in I^k.$$

Step 5. Calculate the quantity $\psi(x^{k+1}, \omega^{k+1})$. If $\psi(x^{k+1}, \omega^{k+1}) > 0$, then put $k = k + 1$ and pass to step 2, if $\psi(x^{k+1}, \omega^{k+1}) = 0$, then put $x^* = x^{k+1}$, $\omega^* = \omega^{k+1}$ and stop.

We will investigate the algorithm. It follows from (9) that the row-vector \tilde{p} whose i -th coordinate is defined by the relation

$$\tilde{p}_i = \begin{cases} \frac{a_i \tilde{x} - b_i}{\tilde{\omega} - c \tilde{x}} \geq 0, & i \in I, \\ 0, & i \in \bar{I}, \end{cases} \quad (15)$$

is a permissible vector of the dual problem (2). If the vectors a_i , $i \in I$, are linearly-independent, then (12) and (15) imply that

$$\tilde{p}_i = \frac{a_i \tilde{\xi}}{1 - c \tilde{\xi}}, \quad i \in I. \quad (16)$$

In this case all the $a_i \tilde{\xi} > 0$, $i \in I$, therefore for $\rho < 0$ all the discrepancies $[a_i(\tilde{x} + \rho \tilde{\xi}) - b_i]$, $i \in I$, decrease. It follows from (14) that $\tilde{\rho} = -(\tilde{\omega} - c \tilde{x}) / (1 - c \tilde{\xi})$ minimizes the right side

of (14), but (16) implies that $\bar{\rho} = -(a_i \bar{x} - b_i) / a_i \bar{x}$, all the discrepancies $[a_i(\bar{x} + \bar{\rho} \bar{x}) - b_i]$ vanishing. Therefore, if at some stage the rows a_i , $i \in I$, are linearly independent, then the step $\bar{\rho}$ is realized and the new value $x = \bar{x} + \bar{\rho} \bar{x}$ is the required optimal point x^* . This obviously corresponds to the last step of the operation of the algorithm.

We prove the convergence of the algorithm. It follows from the description that is operation consists of the sequential choice of points $\omega^0 > \dots > \omega^k > \dots$. On the segment $[\omega^*, \omega^0]$ the function $f(\omega)$ is non-negative, continuous and monotonically increasing, therefore the limit

$$\lim_{k \rightarrow \infty} f(\omega^k) = f(\bar{\omega}) \geq 0.$$

exists. If $f(\bar{\omega}) = 0$, then $\bar{\omega} = \max\{\omega | f(\omega) = 0\} = \omega^*$. The attainment of the point ω^* after a finite number of iterations of the algorithm is guaranteed in this case by the fact that, as demonstrated, the last step of the operation of the algorithm is finite. We establish that the case $f(\bar{\omega}) > 0$ is impossible. From the continuity of the trajectory of local minima it follows that the point

$$\bar{x} = \lim_{k \rightarrow \infty} x(\omega^k)$$

lies on the trajectory and corresponds to the value $\bar{\omega}$, that is, $\bar{x} = x(\bar{\omega})$. The point \bar{x} must belong to at least one of the hyperplanes of the constraints of problem (1), otherwise it would be within the segment $[x^k, x^{k+1}]$, corresponding to some k -th iteration of the algorithm, and the relation $\omega^k > \bar{\omega} > \omega^{k+1}$, would hold, which contradicts the condition $\omega^k \geq \omega$ for any k . For $f(\omega) > 0$ at the point \bar{x} at least one of the constraints of problem (1) is violated, and at the next step of the algorithm some $\omega^h < \bar{\omega}$, will be chosen, which contradicts the definition of the point $\bar{\omega}$ as the limit of a sequence of points $\omega^k \geq \bar{\omega}$ for any k .

Remark 2. Everywhere above it has been assumed that the matrices on the left sides of the systems of linear equations (8) and (12) are non-singular. If at some iteration k this matrix is singular, then in the convex polyhedral set in which the set I^k of constraints of problem (1) is violated, and only one, the trajectory of the local minima assumes the form of a cone. It is obvious that the solution of problem (1) in this case is non-unique. The calculations must be interrupted and the relative disposition of the hyperplanes of the functional and the set I^k investigated.

Remark 3. A version of the algorithm is possible in which x^{k+1} is calculated by the formula $x^{k+1} = x^k + \rho^k \bar{x}^k$, where $\rho^k = -(\omega^k - c x^k) / (1 - c \bar{x}^k)$, minimizes the right side in (14). With this choice of step the point x^{k+1} may not lie on the trajectory of local minima. To calculate $x(\omega^{k+1})$, where $\omega^{k+1} = \omega^k + \rho^k$, it is necessary to calculate the minimum of the function $\psi(x, \omega^{k+1})$. Then x^{k+1} is used as the initial approximation.

Translated by J. Berry.

REFERENCES

1. IVANILOV, Yu. P. Two algorithms for the solution of the concave programming problem. In: *Theory of optimal decisions* (Teoriya optimal'nykh reshenii), No. 4, 13-21, IK Akad. Nauk Ukr SSR, Kiev, 1969.
2. IVANILOV, Yu. P. and PETROV, A. A. Some methods of solving problems of optimal planning for dynamic models of production. In: *Cybernetics - in the service of communism!* (Kibernetika - na sluzhbu kommunizmu!), No. 6, 51-64, "Energiya", Moscow, 1971.

3. LEBEDEV, V. Yu. An approximate algorithm for the solution of the linear programming problem. *Zh. vychisl. Mat. mat. Fiz.*, 14, 4, 1052-1058, 1974.
4. FIACCO, A. and McCORMICK, G. *Non-linear programming. Sequential unconstrained minimization techniques* (Nelineinoe programmirovaniye. Metody posledovatel'noi bezuslovnoi minimizatsii), "Mir", Moscow, 1972.

THE STABILITY AND ASYMPTOTIC ESTIMATION OF THE SOLUTION OF THE INVERSE PROBLEM WITH A SMALL PARAMETER*

I. V. SIMONOV

Moscow

(Received 14 May 1975)

THE TOPICS of the stability with respect to a parameter of the solution of a class of inverse problems, and the estimation of the error of the solutions of these problems are investigated. Theorems are proved which reduce the problems posed to similar direct problems.

We first consider a physical problem of an illustrative nature. Suppose a rigid infinite plate is incident normally at high velocity on the surface of a half-space occupied by a solid strongly porous medium. The picture of the resulting plane one-dimensional motion of the medium is as follows: into the interior of the medium a strong shock wave propagates, behind the front of the wave the matter is in a state of relaxation. The behaviour of the medium will be described by the non-linear equation of a shock adiabatic and linearly-elastic relaxation. Then the problem of calculating the motion of the medium admits of the following equivalent mathematical description. The pressure p , the mass velocity u , and the specific volume v in the domain $0 < x < x_0(t)$, $t > 0$, of the Lagrangian variables x , t must satisfy the equations of motion, of continuity and of linearly-elastic relaxation

$$\frac{\partial u}{\partial t} = -\frac{\partial p}{\partial x}, \quad \frac{\partial u}{\partial x} = \frac{\partial v}{\partial t}, \quad \epsilon(p_1 - p) = a^2 v_1^{-2}(v - v_1)$$

and the boundary conditions: at the plate-medium surface of separation for $x=0$, $t>0$ in the form of the equation of motion of the plate

$$m_0 \frac{\partial u}{\partial t} = -p, \quad u(0, 0) = u_0;$$

at the shock wave front for $x=x_0(t)$, $t>0$ ($x_0(t)$ is the unknown boundary of the domain), in the form of the equation of the shock adiabatic and the general relation at the discontinuity

$$p = p_*(v, v_0), \quad \dot{x}_0 = v(v_0 \dot{x}_0 - u), \quad p = \dot{x}_0 u, \quad x_0(0) = 0.$$

Here $p_1 = p_1(x)$ and $v_1 = v_1(x)$ are the local extremal values of p and v attained at the front, $\epsilon = a^2/c^2$, where $a = \dot{x}_0(0)$ (the dot indicates the time derivative), c is the constant speed of sound at the front, m_0 , u_0 are the density of a unit of the surface of the plate and its initial velocity, v_0 is the initial value of the specific volume.

Putting $\epsilon = 0$, it is possible to obtain the same problem for a medium with a stiff relief. We note that for $\epsilon \neq 0$ this system of equations is of hyperbolic type, for $\epsilon = 0$ it is elliptic.

We introduce the function

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 263-267, 1977.

$$q(x) = \lim_{t \rightarrow \infty} v(x, t),$$

describing the remaining distribution of the specific volume and recorded in experiments. We suppose that this limit exists. We will call the problems of determining q for a given p_* for $\epsilon \neq 0$ and $\epsilon = 0$, problems 1 and 2 respectively.

Problem 1 can be solved by a difference method. The equations of problem 2 are integrated, and q is determined as the solution of the equation [1]

$$p_*(q, v_0) = m_0^2 u_0^2 / (v_0 - q)(x + m_0)^2.$$

The solution of the inverse problem of the determination of p_* for a known q (problems 3, 4, for $\epsilon \neq 0$ and $\epsilon = 0$, respectively) is physically interesting. These are typical problems of determining a function of state by indirect measurements. The solution of problem 4 is given in implicit form [1]:

$$v = q(x), \quad p = m_0^2 u_0^2 / (v_0 - q)(x + m_0)^2 \equiv F(q, x).$$

The algorithm for solving problem 3 is unknown. Nevertheless, it is desirable to estimate the error of the solution of problem 4 due to simplifying the model of the medium. Below a mathematical investigation of the situation discussed is presented. A stability theorem is proved in which it is stated that for the stability of the solution of some class of inverse problems with respect to a parameter the uniform stability of the solution of the corresponding direct problems with respect to the same parameter is sufficient. The estimate of the error of the solution of the converse problem is obtained in terms of the norm of the error of the solution of the direct problem. Thereby the questions of the stability and estimation in the converse problem are reduced to the similar problems in the direct problem.

1. We will start from the assumption that there exist several perturbed and unperturbed problems of determining a function q for a specified function p , and also the problems converse to them of determining p for q .

Let A and A_0 be perturbed and unperturbed operators (in general non-linear) transforming p into q , specified on some set P , and Q_ϵ and Q_0 be the domains of values of these operators respectively, where $P \subset E_1$, $Q_\epsilon \subset E_2$, $Q_0 \subset E_2$, and E_1 and E_2 are two linear normed spaces of functions.

The perturbed operator A_ϵ depends on the parameter ϵ and is defined in the half-interval $\mathcal{E} = (0, \epsilon_0]$. The operator A is obtained if in A_ϵ we formally put $\epsilon = 0$. We denote by β the difference

$$\beta(\epsilon, p) = A_\epsilon p - A_0 p \quad (1.1)$$

and we will say that the operator A_0 approximates the operator A_ϵ on the set P , if for any function $p \in P$ the following condition is satisfied:

$$\beta(\epsilon, p) \rightarrow 0 \quad \text{as} \quad \epsilon \rightarrow 0. \quad (1.2)$$

Here and overleaf convergence is understood in the strong sense as convergence in norm.

We suppose that each of the mappings A_ϵ and A_0 is one-to-one, and let A_ϵ^{-1} and A_0^{-1} be inverse operators transforming q into p . We also suppose that the set $Q = Q_\epsilon \cap Q_0$ — the intersection of Q_0 and the whole family of sets $\{Q_\epsilon, \epsilon \in \mathcal{E}\}$ — is not empty, and we will use the notation

$$p_\epsilon = A_\epsilon^{-1}q, \quad p_0 = A_0^{-1}q, \quad \alpha(\epsilon, q) = p_\epsilon - p_0, \quad q \in Q.$$

Therefore, if by means of each of the four operators introduced $A_\epsilon, A_0, A_\epsilon^{-1}, A_0^{-1}$ the solution of one of the four problems is constructed (the direct perturbed, and unperturbed and the inverse perturbed and unperturbed, respectively), then the quantities α and β have the meaning of the errors of the solutions of the unperturbed problems, direct and inverse respectively. The approximation condition (1.2) denotes stability of the solution of the direct problem with respect to the parameter ϵ .

We prove the following stability theorem.

Theorem 1

Let the operator A_0^{-1} be continuous on Q_0 and $\beta(\epsilon, p)$ tend uniformly to zero on the set P as $\epsilon \rightarrow 0$. Then the operator A_0^{-1} approximates the perturbed operator A_ϵ^{-1} on the set Q as $\epsilon \rightarrow 0$.

Proof. To establish the connection between the quantities β and α we arrive at the following operator equation. By Eq. (1.1), $A_\epsilon p_\epsilon - A_0 p_\epsilon = \beta(\epsilon, p_\epsilon)$.

From this, by mutual uniqueness $A_\epsilon p_\epsilon = q, A_0 p_\epsilon = q - \beta(\epsilon, p_\epsilon)$.

Multiplying both sides of the last equation by A_0^{-1} , we obtain

$$p_\epsilon = A_0^{-1}(q - \beta), \quad \alpha(\epsilon, q) = A_0^{-1}[q - \beta(\epsilon, p_\epsilon)] - A_0^{-1}q. \quad (1.3)$$

Equations (1.3) imply that if the operator A_0^{-1} is continuous and

$$\beta(\epsilon, p_\epsilon) \rightarrow 0 \quad \text{as} \quad \epsilon \rightarrow 0, \quad (1.4)$$

then the statement of the theorem will be valid

$$\alpha(\epsilon, q) \rightarrow 0 \quad \text{as} \quad \epsilon \rightarrow 0, \quad q \in Q.$$

The condition of continuity of the operator A_0^{-1} is stipulated in the formulation of Theorem 1. The validity of (1.4) simply follows from the condition of uniform convergence of β .

In the theorem it is asserted that for the stability of the solution of the inverse problem with respect to a parameter the uniform stability of the solution of the direct problem with respect to the same parameter is sufficient.

We note that the condition that β tend uniformly to zero on the whole set P is too strong a requirement. Indeed, to prove the assertion (1.4) for each fixed $q \in Q$ it is sufficient that $\beta(\epsilon, p) \rightarrow 0$ as $\epsilon \rightarrow 0$ uniformly only on the one-parameter family of functions $\{p_\epsilon = A_\epsilon^{-1}q, \epsilon \in \mathcal{E}\}$.

2. If the operator A_0^{-1} is Frechét differentiable on the set Q_0 , then the right side of (1.3) can be represented in the form

$$A_0^{-1}(q-\beta) - A_0^{-1}q = -\beta A'q + \omega(q, \beta), \quad (2.1)$$

where A' is a linear bounded operator,

$$\|A'q\|_{E_1} \leq C\|q\|_{E_2}, \quad q \in Q_0,$$

and $\omega(q, \beta)$ satisfies the condition

$$\lim_{\|\beta\| \rightarrow 0} (\|\omega(q, \beta)\|_{E_1} \|\beta\|_{E_2}^{-1}) = 0.$$

The following theorem on estimation holds.

Theorem 2

Let the conditions of Theorem 1 be satisfied, and also let the operator A_0^{-1} be Frechét differentiable on the set Q_0 and let the following conditions be satisfied:

$$1) \|\beta(e, p)\|_{E_1} \neq 0, \quad p \in P, \quad e \in \mathcal{E};$$

2) for any fixed function $p \in P$ and any sequence of functions $\{p_n\} \in P$, $n=1, 2, \dots$, such that $p_n \rightarrow p$ as $n \rightarrow \infty$, the sequence $\{\|\beta(e, p_n)\| \|\beta(e, p)\|^{-1}\}$ converges uniformly on the set \mathcal{E} to unity as $n \rightarrow \infty$.

Then as $\varepsilon \rightarrow 0$ the following asymptotic estimate holds ($q \in Q$, $p_0 = A_0^{-1}q$):

$$\|\alpha(e, q)\|_{E_1} \leq C_1 \|\beta(e, p_0)\|_{E_1}, \quad C_1 = C + o(1).$$

Proof. From (1.3), (2.1) taking into account (1.4) we have

$$\|\alpha(e, q)\| \leq C \|\beta(e, p_e)\| + \|\omega[q, \beta(e, p_e)]\|. \quad (2.2)$$

We prove the validity of the representation

$$\|\beta(e, p_e)\| = \|\beta(e, p_0)\| + \omega_1(e, q), \quad (2.3)$$

where ω_1 satisfies the condition

$$\lim_{\varepsilon \rightarrow 0} [\omega_1(e, q) \|\beta(e, p_0)\|^{-1}] = 0.$$

It is required to prove that for any $q \in Q$

$$\lim_{\varepsilon \rightarrow 0} [\|\beta(e, p_e)\| \|\beta(e, p_0)\|^{-1}] = 1. \quad (2.4)$$

Let $\{e_n\}$, $n=1, 2, \dots$, be some arbitrary infinitely small sequence:

$$\lim_{n \rightarrow \infty} e_n = 0.$$

Then, by the stability condition assumed

$$\lim_{n \rightarrow \infty} p_{e_n} = p_0.$$

By condition 2) of Theorem 2, $\|\beta(\epsilon, p_{\epsilon_n})\| \|\beta(\epsilon, p_0)\|^{-1} \rightarrow 1$ as $n \rightarrow \infty$ uniformly on the whole set \mathcal{E} , and this means that for any $\delta > 0$ and any $\epsilon \in \mathcal{E}$ there exists a natural number $m = m(\delta)$ such that for any $n > m$ the following inequality is satisfied:

$$|\|\beta(\epsilon, p_{\epsilon_n})\| \|\beta(\epsilon, p_0)\|^{-1} - 1| < \delta. \quad (2.5)$$

In place of ϵ in the inequality (2.5) it is possible to substitute ϵ_n , and it is then proved that for any infinitely small sequence $\{\epsilon_n\}$ and any function $q \in Q$

$$\lim_{n \rightarrow \infty} (\|\beta(\epsilon_n, p_{\epsilon_n})\| \|\beta(\epsilon_n, p_0)\|^{-1}) = 1.$$

This proves the validity of (2.4) and (2.3).

Using (2.3) and (2.2), we obtain

$$\|\alpha(\epsilon, q)\| \leq (C + \gamma(\epsilon, q)) \|\beta(\epsilon, p_0)\|, \quad \gamma = (\|\omega\| + \omega_1) \|\beta\|^{-1} \rightarrow 0 \text{ as } \epsilon \rightarrow 0, \quad (2.6)$$

which is what it was required to prove.

3. The resulting obtained permit us to judge the stability of the solution of the inverse problem with a parameter and to estimate this solution by an indirect method without having recourse to the solution of the inverse perturbed problem or the corresponding linearized problem of small perturbations, as is done traditionally. For an estimate by formula (2.6) it is necessary to obtain the solution of the inverse unperturbed problem and estimate the error of the solution of the corresponding direct problem and the norm of the operator A_0^{-1} .

We must mention the strong constraints imposed on the properties of the operators in the proofs of the theorems. The problem considered in the introduction is an example where the conditions of the theorem are satisfied. Of course, not all of them are mathematically strictly verified: some of the conditions, such as the requirement of uniform convergence of β and non-emptiness of the set Q , are checked by a numerical experiment [2], that is, on a discrete set of functions and points. The continuity of the inverse operator A_0^{-1} , which is obviously the strongest requirement, in the case of some minimal and natural constraints on the set of functions p_* and q follows from the continuity of the linear-fractional function. We note that this condition can be replaced by the requirement of compactness of the set P [3].

The results of paragraphs 1, 2 imply an estimate of the difference of the solutions of problems 3, 4 in the form

$$p_* - p_0 \approx \frac{\partial F}{\partial q} \bigg|_{q=q(x)} [q_*(x) - q(x)],$$

or

$$p_{*s} - p_{*0} \approx \frac{\partial F}{\partial x} \bigg|_{x=x(q)} [x_*(q) - x(q)],$$

where $q_\epsilon(x)$ is the solution of problem 1 for $p_* = p_{*0}$ and $x(q)$ ($x_\epsilon(q)$) is the function inverse to $q(x)$ ($q_\epsilon(x)$).

Translated by J. Berry

REFERENCES

1. SIMONOV, I. V. The inverse problem of determining the shock adiabat of a strongly deformed medium and the estimation of its solution. *Izv. Akad. Nauk SSSR. Mekhan. tverdogo tela*, No. 4, 122-128, 1976.
2. SIMONOV, I. V. Estimation of the solution of a problem of the propagation of waves in a medium with strict relief. *Izv. Akad. Nauk SSSR. Mekhan. tverdogo tela*, No. 6, 120-125, 1974.
3. TIKHONOV, A. N. and ARSENIN, V. Ya. *Methods of solving ill-posed problems* (Metody resheniya nekorrektnykh zadach), "Nauka", Moscow, 1974.

A NUMERICAL METHOD OF SOLVING THREE-DIMENSIONAL DIFFRACTION PROBLEMS*

A. L. GAPONENKO

Moscow

(Received 26 November 1974; revised 7 March 1975)

A MODIFICATION of the method of non-orthogonal series is proposed for solving three-dimensional diffraction problems. A justification of the method is given and numerical results are presented.

Recently methods of solving three-dimensional diffraction problems with axial symmetry have been greatly developed. The presence of axial symmetry makes the problem essentially two-dimensional, which facilitates the search for the solution of problems of this type. One possible method of solving three-dimensional diffraction problems is the method of non-orthogonal series. However, it requires the summation of a large number of terms of the series, which considerably lengthens the calculations and leads to loss of accuracy. The modification of the method of non-orthogonal series proposed in this paper permits the accuracy to be increased for a fixed number of basis functions.

1. Statement of the problem

We consider the three-dimensional problem of the diffraction of a scalar field by a body of arbitrary shape. Let a plane wave $u_0 = C_0 \exp(ikz)$ be incident on a bounded body V with boundary S . We suppose that the boundary S is a surface of Lyapunov type. The total field U can be represented in the form

$$U = u_0 + u,$$

where u_0 and u are the incident and diffracted field respectively. We consider the Dirichlet problem. Then the diffracted field will be the solution of the following problem:

$$\begin{aligned} \Delta u(M) + k^2 u(M) &= 0, & M \in V, \\ u|_S &= -C_0 \exp(ikz)|_S, \\ du/dr - iku &= o(1/r) \text{ as } r \rightarrow \infty, \end{aligned} \quad (1)$$

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 267-272, 1977.

where V_e is the domain exterior in relation to the surface S , and $k=\text{const}$ is the characteristic of the domain V_e .

It is known, with the assumptions indicated above, that problem (1) has a unique solution $u(M)$ [1].

2. The modified method of non-orthogonal series

We consider an arbitrary closed surface S_1 , situated entirely inside the body V . We suppose that the surface S_1 is not resonant, that is, that the interior homogeneous Dirichlet problem for the equation $\Delta u + k^2 u = 0$ has only a trivial solution. We consider the system of functions $\psi_n(M)$, proposed by V. D. Kupradze [2]:

$$\psi_n(M) = \frac{\exp[ikR(M, M_n)]}{R(M, M_n)},$$

where the points M_n form a countable, everywhere dense set on the surface S_1 , and $R(M, M_n)$ is the distance between the points M and M_n . Using the method proposed in [2] we prove the following theorem.

Theorem 1

The system of functions $\{\psi_n(M)\}$ is linearly independent and complete in the space $L_2(S)$.

Proof. We prove the linear independence of the functions $\psi_n(M)$, $n=1, 2, \dots$. Let the system $\{\psi_n(M)\}$ be linearly dependent, that is, let N and C_i , $i=1, 2, \dots, N$, where $|C_1| + \dots + |C_N| \neq 0$, exist such that

$$\sum_{i=1}^N C_i \psi_i(M) = 0, \quad M \in S.$$

We consider the function

$$W(M) = \sum_{i=1}^N C_i \psi_i(M), \quad M \in \bar{V}_e.$$

It is easy to verify that the function $W(M)$ equals zero everywhere outside the surface S_1 . But in a sufficiently small neighborhood of the point M_j the function ψ_j assumes a value arbitrarily great in absolute magnitude, while the other terms in the expression for $W(M)$ are bounded. Hence we obtain $C_j=0$, $j=1, 2, \dots, N$, which contradicts the assumption about the coefficients C_j . Consequently, the system $\{\psi_n(M)\}$ is linearly independent.

To prove completeness it is sufficient to show that for any $\varepsilon > 0$ and for any function $\alpha(M) \in L_2(S)$ we can find N and C_i , $i=1, 2, \dots, N$, such that

$$\left\| \alpha(M) - \sum_{i=1}^N C_i \psi_i(M) \right\|_{L_2(S)} < \varepsilon.$$

This statement is equivalent to the following: if $\alpha(M) \in L_2(S)$ and

$$\int_S \alpha(M) \psi_n(M) dS = 0, \quad n=1, 2, \dots, \quad (2)$$

then $\alpha(M) = 0$ almost everywhere on S . We first prove (2) for a continuous function $\beta(M)$. Let

$$\int_S \beta(M) \psi_n(M) dS = 0.$$

We consider the function

$$F(P) = \int_S \frac{\beta(M) \exp[ikR(M, P)]}{R(M, P)} dS_M.$$

It is easy to show that $F(P) = 0$ in all space. This implies that

$$(\partial F / \partial n)_{\text{outward}} = (\partial F / \partial n)_{\text{inward}} = 0.$$

However, by the property of the potential of a simple layer

$$(\partial F / \partial n)_{\text{outward}} - (\partial F / \partial n)_{\text{inward}} = 4\pi\beta(M).$$

Consequently, $\beta(M) = 0$ on S . We have proved the closedness of $\{\psi_n(M)\}$ in $C(S)$ in the sense of the metric $L_2(S)$, consequently, we have proved the completeness of $\{\psi_n(M)\}$ in the sense of the metric $L_2(S)$. Since the set of continuous functions is everywhere dense in $L_2(S)$, we have thereby proved the completeness of the system $\{\psi_n\}$ in the space $L_2(S)$. Theorem 1 is proved completely.

Theorem 2

If the sequence $\{u^n(M)\}$ possesses the following properties:

1) $u^n(M)$ satisfies in V_e the Helmholtz equation and the radiation condition at infinity, $n = 1, 2, \dots$,

2) $\|u^n(\bar{M}) + C_0 \exp(ikz)\|_{L_2(S)} = \delta_n \rightarrow 0$ as $n \rightarrow \infty$,

and if V_e' is an arbitrary closed domain, $V_e' \subset V_e$, and u is the exact solution of problem (1), then

$$\|u - u^n\|_{C(V_e')} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Proof. We construct the difference between the exact and the approximate solution $w(M) = u - u^n$. Then

$$w(M) = \int_S \frac{\partial G}{\partial n_P}(M, P) w(P) dS_P, \quad (3)$$

where $G(M, P)$ is Green's function of the exterior Dirichlet problem for the Helmholtz equation. Since in any closed region V_e' , situated within the domain V_e , $\partial G(M, P) / \partial n_P$ is bounded, then,

applying to (3) the Cauchy inequality, we obtain

$$\|w(M)\|_{C(V_e')} < K\delta_n \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Here K is a constant estimating the integral

$$\int_S \left| \frac{\partial G}{\partial n_P}(M, P) \right|^2 dS_P < K^2, \quad M \in V_e'.$$

Theorem 2 is completely proved.

Let $\{\delta_n\}$ be an arbitrary numerical sequence, $\delta_n \rightarrow 0$ as $n \rightarrow \infty$. By Theorem 1, for any n we can find a number $N = N(n)$, points $\{M_i^n\}$, $M_i^n \in S_1$, $i=1, 2, \dots, N$, inducing corresponding functions $\psi_i^n(M)$, $i=1, 2, \dots, N$, and also coefficients $\{C_i^n\}$, $i=1, 2, \dots, N$, such that the inequality

$$\left\| \sum_{i=1}^N C_i^n \psi_i^n(M) + C_0 \exp(ikz) \right\|_{L_2(S)} < \delta_n.$$

will be satisfied. We write

$$u_N^n(M) = \sum_{i=1}^N C_i^n \psi_i^n(M). \quad (4)$$

Since the sequence (4) defined above satisfies the conditions of Theorem 2, it satisfies

$$\|u - u_N^n\|_{C(V_e')} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Therefore the elements of the sequence (4) can be regarded as an approximate solution of problem (1). Therefore, to find an approximate solution of the original diffraction problem it is sufficient to solve the problem of minimizing the functional

$$\left\| \sum_{i=1}^N C_i \psi_i(M) + C_0 \exp(ikz) \right\|_{L_2(S)}. \quad (5)$$

Usually a functional of the type (5) is minimized by searching for coefficients of an expansion for a fixed system of basis functions. In the modified method of non-orthogonal series it is proposed to minimize the functional (5) not only allowing for the variation of the coefficients C_i , but also allowing for the variation of the basis functions themselves. In our case the basis functions are varied by varying the coordinates of the points $M_i(x_i, y_i, z_i)$. For a fixed number of basis functions this procedure permits us to find the optimal distribution of the points $\{M_i\}$ for a given perturbation. If we note that the function $\exp[ikR(M, M_j)]/R(M, M_j)$ is a scalar field created by a point source situated at the point M_j , then the procedure indicated can be given a certain physical interpretation. We seek a disposition of point sources within the body and a distribution of their intensities such that the total field of these sources will be the same as the diffraction of a plane wave.

It is also possible to give a mathematical interpretation of the algorithm for searching for the coefficients of the expansion C_i and the simultaneous search for the coordinates of the

points M_j . In the solution of the problem of approximating a given function u_0 by a segment

of the series $\sum_{n=1}^N C_n \psi_n$, where $\{\psi_n\}$ is a complete system of functions, the following situation may arise. The hyperplane formed by the elements ψ_n , $n=1, 2, \dots, N$, will be almost orthogonal to the element u_0 . It is impossible to obtain a better approximation to the element u_0 by means of the given N functions. In the modified method of non-orthogonal series we seek N functions $\{\psi_n\}$, such that the norm of the projection of the element u_0 on the hyperplane formed by the given N elements $\{\psi_n\}$, is a maximum.

In the solution of diffraction problems the final aim of the investigation is often to obtain radiation patterns. We introduce a system of spherical coordinates with centre at the point $O \in V$. Let (r, θ, φ) be the coordinates of the point M . Then the field $u(M)$ in the far zone can be written in the form

$$u(M) = \frac{\exp(ikr)}{r} D(\theta, \varphi) + o\left(\frac{1}{r}\right),$$

where $D(\theta, \varphi)$ is the field radiation pattern. In the case where an approximate solution

$$u_N(M) = \sum_{i=1}^N C_i \psi_i(M), \quad (6)$$

is known, the corresponding approximation to the actual radiation pattern has the form

$$D_N(\theta, \varphi) = \sum_{n=1}^N C_n \exp[-ik(\sin \theta \cos \varphi x_n + \sin \theta \sin \varphi y_n + \cos \theta z_n)].$$

Here x_n, y_n, z_n are the Cartesian coordinates of the point M_n .

Remark 1. The case of the perturbation of a plane wave was considered only for definiteness. The description of the method is unchanged if a perturbation of any other form is considered.

Remark 2. The method transfers without difficulty to the case of the second and other boundary value problems.

3. Numerical realization of the method

Thus, the original problem (1) has been reduced to the problem of finding the minimum of a function of $5N$ variables:

$$F(\operatorname{Re} C_j, \operatorname{Im} C_j, x_j, y_j, z_j) = \iint_S \left| \sum_{j=1}^N C_j \frac{\exp[ikR(M, M_j)]}{R(M, M_j)} + C_0 \exp(ikz) \right|^2 dS_M.$$

Various methods of minimizing functions of many variables can be used to solve this problem. For the numerical realization the method of conjugate gradients (the Fletcher-Reeves method [3]) was used. This method was chosen as the most appropriate to the specific nature of the

given problem. To illustrate the operation of the algorithm we give the results of a calculation of the following two simulated cases.

Case 1. On an ellipsoid with axes, $a = 1$, $b = 2$, $c = 3$ there is incident from the direction of the greatest axis a plane wave, with wave number $k = 0.6$, the amplitude of the incident wave $C_0 = 10$.

Case 2. On a sphere of radius $r = 1$ there is incident a plane wave,

$$k = 1, C_0 = 10.$$

The calculations were performed on the BESM-6 computer for twenty basis functions. In the case of diffraction by the sphere after 65 steps the Fletcher-Reeves method succeeded in attaining a decrease of the function (5) from 417 to 0.44. The value of the function (5) can be interpreted as the discrepancy in $L_2(S)$. The field at points of the surface S was calculated by Eq. (6). The maximum value of $|u_N + u_0|^2$ can be interpreted as the discrepancy in C . If the value of the discrepancy is referred to the square of the amplitude of the incident field, then we obtain a percentage expression. In the solution obtained for the sphere the values of the discrepancies in L_2 and C were 0.44 and 1.4% respectively. The results obtained for diffraction by the sphere were the same as those published in [4].

In the case of diffraction by the ellipsoid after 30 steps the function (5) was reduced from 33.18 to 0.39. The values of the discrepancies in L_2 and C amounted to 0.39 and 2.48% respectively. From the data given there is obviously a difference in the number of steps, and consequently also in the solution time of the problem for the sphere and for the ellipsoid. This is explained by the fact that for the ellipsoid a more favourable zero approximation was chosen, taking into account some of the regularities in the solution of the diffraction problem for the sphere. It must be mentioned that the function (5) is not convex, so that there apparently exists a set of local minima, among which one or more are global. However, for the solution of the problem it is unimportant whether we begin in the neighborhood of a global or in the neighborhood of a local minimum, since the accuracy of the solution depends only on the value of the discrepancy in $L_2(s)$.

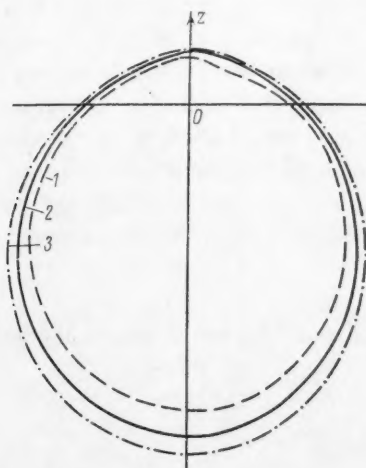


FIG. 1.

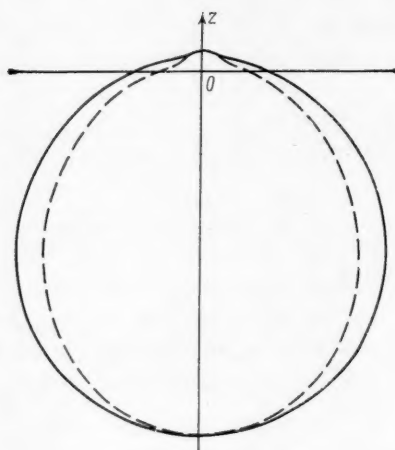


FIG. 2.

Figure 1 shows the radiation pattern of the field diffracted by the sphere after 25 (curve 1), 35 (curve 2) and 65 (curve 3) steps with values of the discrepancies in L_2 3.3, 1.08 and 0.44% respectively. It should be mentioned that in the solution of the problem of diffraction by the sphere (in this case the exact solution is symmetrical with respect to the angle φ) a sufficiently good symmetry with respect to the angle φ was obtained. The radiation pattern obtained deviates from symmetry by not more than 0.16%.

Figure 2 shows the radiation pattern of the field diffracted by the ellipsoid. The sections of the radiation pattern are shown in the OXZ plane (continuous line) and the OYZ plane (dashed line).

In the process of solving the above two problems, and also a number of other problems of the diffraction of a plane wave by three-dimensional bodies, some tendencies in the behaviour of the coefficients and in the disposition of the internal sources could be distinguished. The internal sources are so arranged that relative to the geometrical centre of the body on which diffraction occurs, they are displaced toward the front of the incident wave. For the sources furthest from the wave front $\text{Re}C_i$ and $\text{Im}C_i$ are negative, at those close to the wave front they are positive. As the sources approach the wave front, $\text{Re}C_i$ and $\text{Im}C_i$ increase monotonically. If we choose a zeroth approximation allowing for the given behaviour, the process of convergence of the method can be improved.

The author thanks V. V. Kravtsov for his guidance and interest.

Translated by J. Berry.

REFERENCES

1. KUPRADZE, V. D. *Boundary value problems of the theory of oscillations and integral equations* (Granichnye zadachi teorii kolebaniy i integral'nye uravneniya). Gostekhizdat, Moscow-Leningrad, 1950.
2. KUPRADZE, V. D. On the approximate solution of problems of mathematical physics. *Usp. mat. Nauk*, 22, 2, (134), 59-107, 1967.
3. FIACCO, A. and McCORMICK, G. *Non-linear programming: Sequential unconstrained minimization techniques* (Nelineinoe programmirovaniye. Metody posledovatel'noi besulovnoi minimizatsii), "Mir", Moscow, 1972.
4. ASVESTAS, J. S. et al. *Electromagnetic and acoustic scattering by simple shapes*. North-Holland Publ. Co., Amsterdam, 1969.

SOLUTION OF THE INVERSE PROBLEM OF THE DISPERSION OF AN ELECTROMAGNETIC PULSE IN A CONDUCTING MEDIUM*

V. V. YANKOV

Moscow

(Received 3 March 1975; revised 27 June 1975)

A SIMPLE analytic expression is obtained for the initial shape of a plane electromagnetic wave pulse in terms of the shape acquired by the pulse during propagation in a homogeneous conducting medium, in the form of the exact solution of an integral equation of the first kind of the convolution type.

It is well-known that the propagation of an electromagnetic perturbation in a homogeneous isotropic conducting medium is described by a partial differential equation of the hyperbolic type of the form of the "telegraph equation"

$$\Delta u = \frac{\epsilon\mu}{c^2} \frac{\partial^2 u}{\partial t^2} + \frac{4\pi\sigma\mu}{c^2} \frac{\partial u}{\partial t},$$

where u is any of the components of the electromagnetic field E_x, E_y, E_z or H_x, H_y, H_z in an arbitrary Cartesian coordinate system, ϵ and μ are the dielectric constant and magnetic permeability independent of frequency, σ is the constant electrical conductivity of the medium, and c is the velocity of light in a vacuum (see, for example, [1, 2]). Here both in the absence of dispersion of the dielectric constant, and also in the frequently encountered case where it is permissible to neglect the effect on the spreading of the pulse of displacement currents in comparison with the conduction current (see, for example, [3]) the change in shape of the pulse $u(x, y, z, t)$ (without allowing for the time delay of the pulse as a whole) satisfies the parabolic heat-conduction equation

$$\Delta u = \frac{1}{a^2} \frac{\partial u}{\partial t}, \quad a^2 = \frac{c^2}{4\pi\sigma\mu}.$$

Its solution $u(x, t)$ in the simplest, plane case

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{a^2} \frac{\partial u}{\partial t} \quad (1)$$

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 273-276, 1977.

of a one-dimensional boundary value problem for a half-bounded space $x > 0$ without initial conditions $t > -\infty$ and with the boundary condition

$$u(0, t) = f(t) \quad (2)$$

for any bounded piecewise-continuous function $f(t)$ has the form (see, for example, [2])

$$u(x, t) = \int_{-\infty}^t f(\tau) K(x, t-\tau) d\tau, \quad (3)$$

where the binary source function

$$K(x, t) = \begin{cases} \frac{1}{2\pi^{1/2}} \frac{x}{a} t^{-1/2} \exp\left(-\frac{1}{4t} \frac{x^2}{a^2}\right), & t > 0, \\ 0, & t < 0. \end{cases} \quad (4)$$

In the physical problem considered the impulse function (4) describes a deformation proportional to the propagation in the conducting medium of a perturbation, defined as a δ -function $f(t) = \delta(t)$ and actuated in the plane $x = 0$.

We are interested in the problem of determining the original shape of the pulse $f(t)$ from the shape $u(x, t)$ distorted by transient phenomena in a homogeneous transmission line with a pulse characteristic of the form (4). Previously (see, for example, [4, 5]) it was discussed exclusively from the point of view of the possibility of finding by using a computer the solution directly of the convolution type integral equation (3) by one of the general methods for the approximate solution of ill-posed problems — a regularization method.

Meanwhile it will be shown below that the efficient use of some general analytic properties of the solutions of the heat-conduction equation (1) together with the specific integral relation for the kernel (4) of the integral equation.

$$\int_0^t K(x_1, \tau) K(x_2, t-\tau) d\tau = K(x_1+x_2, t) \quad (5)$$

(see, for example, [6]) leads to a characteristic inversion of formula (3), which is in principle equivalent to obtaining the exact solution of the integral equation (3).

Therefore, from the mathematical aspect the determination of the shape of the electromagnetic pulse $f(t)$ at the plane boundary $x = 0$ of the homogeneous half-space $x > 0$ from the shape of the same pulse $u(x, t)$ at a distance $x > 0$ from the boundary belongs, in general, to the category of so-called inverse problems (see, for example, [7]). However, unlike the usual formulation of similar problems, for which their ill-posedness is stipulated at the very beginning and in some way or other is immediately introduced into the scheme for the approximate solution of the integral equation (3), we confine ourselves to only an adequate representation of the solution $f(t)$ of the given inverse problem in the most general analytic form, that is, essentially, to the reduction of one ill-posed problem to another simpler one.

In other words, we here want to find the exact solution of the integral equation (3) on the set F of piecewise-continuous functions $f(t) \in F$ only for such initial functions $u(x, t)$ as necessarily belong to the set $u(x, t) \in AF$, where AF is the image of the set F in the mapping of

the latter performed by the integral operator

$$Af \equiv \int_{-\infty}^t K(x, t-\tau) f(\tau) d\tau.$$

For this purpose we use primarily the fact that the solutions $u(x, t)$ of Eq. (1) are analytic functions of the variable x [8], which can be expanded in an infinite Taylor power series in powers of the difference $x_0 - x$ in the neighborhood of any point $x > 0$:

$$u(x_0, t) = u(x, t) + \frac{x_0 - x}{1!} \frac{\partial u(x, t)}{\partial x} + \frac{(x_0 - x)^2}{2!} \frac{\partial^2 u(x, t)}{\partial x^2} + \dots, \quad (6)$$

if $|x_0 - x| < \infty$. Then, in particular, in the limit as $x_0 \rightarrow 0$ the expansion

$$\lim_{x_0 \rightarrow 0} u(x_0, t) = u(x, t) - \frac{x}{1!} \frac{\partial u(x, t)}{\partial x} + \frac{x^2}{2!} \frac{\partial^2 u(x, t)}{\partial x^2} - \dots,$$

holds, which in consequence of the boundary condition (2), understood in the sense of the limiting value

$$u(0, t) = \lim_{x \rightarrow 0} u(x, t) = f(t),$$

must be identical with the unknown function $f(t)$ at all points of its continuity

$$f(t) = u(x, t) - \frac{x}{1!} \frac{\partial u(x, t)}{\partial x} + \frac{x^2}{2!} \frac{\partial^2 u(x, t)}{\partial x^2} - \dots, \quad (7)$$

since the function (3) as has the limiting value $f(t)$, only if the function $f(t)$ is continuous at the point t (see, for example, [6]). Then, we expand by (6) in a Taylor series with centre at the point x also the auxiliary function $u(2x, t)$, equal on the other hand, by (3), (5), to the convolution of the original function $u(x, t)$ with the impulse function (4), namely

$$\int_{-\infty}^t u(x, \tau) K(x, t-\tau) d\tau = u(x, t) + \frac{x}{1!} \frac{\partial u(x, t)}{\partial x} + \frac{x^2}{2!} \frac{\partial^2 u(x, t)}{\partial x^2} + \dots \quad (8)$$

We then add separately the left and right sides of formulas (7), (8), after which we solve the result for $f(t)$:

$$f(t) = 2 \left[u(x, t) + \sum_{n=1}^{\infty} \frac{x^{2n}}{(2n)!} \frac{\partial^{2n} u(x, t)}{\partial x^{2n}} \right] - \int_{-\infty}^t u(x, \tau) K(x, t-\tau) d\tau. \quad (9)$$

We now consider the general equation for $u(x, t)$

$$\frac{\partial^{2n} u(x, t)}{\partial x^{2n}} = \frac{1}{a^{2n}} \frac{\partial^n u(x, t)}{\partial t^n}, \quad (10)$$

obtained by differentiation $n - 1$ times with respect to time of Eq. (1) with subsequent change of the order of differentiation on the left side with respect to t and with respect to x and taking (1) into account, since the solution (3) of Eq. (1) is analytic in x and has continuous derivatives of all orders with respect to t for $0 < x < \infty$ [8]. Finally, replacing under the summation sign in (9) the spatial derivatives by time derivatives, using (10), we arrive at a simple analytic expression for the solution of the inverse problem posed:

$$f(t) = 2 \left[u(x, t) + \sum_{n=1}^{\infty} \frac{1}{(2n)!} \left(\frac{x}{a} \right)^{2n} \frac{\partial^{2n} u(x, t)}{\partial t^{2n}} \right] - \int_{-\infty}^t u(x, \tau) K(x, t-\tau) d\tau. \quad (11)$$

With the assumptions made about the class of functions $f(t)$, the uniqueness of the solution (11) of the integral equation (3) within the limits of the statement of the problem formulated above, follows directly from the proposed method of sequential derivation of the inverse formula (11).

In particular, formula (11) gives the correct solution of the inverse problem in the case of a monochromatic boundary function of time $f(t) = \text{const} \cdot \exp(i\omega t)$ with an arbitrary frequency ω , when the exact connection between the functions $u(x, t)$ and $f(t) = u(x, t) \exp[(x/a)(i\omega)^{1/2}]$ is known from the solution of the corresponding direct problem in complex form (see, for example, [2])

$$u(x, t) = \text{const} \exp \left[-\frac{x}{a} (i\omega)^{1/2} + i\omega t \right], \quad (12)$$

which is easily verified by substitution of (12) into (11) and summation of the series.

Therefore, the problem of reconstructing the original undistorted shape $f(t)$ of an electromagnetic pulse, distorted in traversing the path x in a homogeneous conducting medium, is rigorously reduced to the operations of differentiation and integration with respect to time of the observed shape $u(x, t)$ of the pulse. In the case of the performance of an approximate calculation of the pulse shape with a given error, the number of terms of the series (11) to be calculated will naturally depend on the value of the expansion parameter x/a and on the time characteristics of the pulse investigated.

Translated by J. Berry.

REFERENCES

1. STRATTON, J. A. *Electromagnetic theory* (Teoriya elektromagnetizma). Gostekhizdat, Moscow-Leningrad, 1948.
2. TIKHONOV, A. N. and SAMARSKII, A. A. *The equations of mathematical physics* (Uravneniya matematicheskoi fiziki). "Nauka", Moscow, 1966.
3. NOVIKOV, V. V. Review of papers on the propagation of impulsive electromagnetic signals in conducting media and over the surface of the earth. In: *Problems of diffraction and wave propagation* (Problemy difraktsii i rasprostraneniya voln), No. II, 7-38, Izd-vo LGU, Leningrad, 1962.

4. ARSENIN, V. Ya. and IVANOV, V. V. The solution of certain convolution type integral equations of the first kind by the regularization method. *Zh. vychisl. Mat. mat. Fiz.*, 8, 2, 310–321, 1968.
5. TIKHONOV, A. N. On the solution of ill-posed problems. *Dokl. Akad. Nauk SSSR*, 151, 3, 501–504, 1963.
6. DEUTSCH, G. *Handbook on the practical application of the Laplace transformation* (Rukovodstvo k prakticheskomu primeneniyu preobrazovaniya Laplasy), Fizmatgiz, Moscow, 1960.
7. TIKHONOV, A. N. and ARSENIN, V. Ya. *Methods of solving ill-posed problems* (Metody resheniya nekorrektnykh zadach), "Nauka", Moscow, 1974.
8. SMIRNOV, V. I. *Course of higher mathematics* (Kurs vysshei matematiki), Vol. IV, Gostekhizdat, Moscow, 1957.

CONVERGENCE OF NEWTON'S ITERATIVE METHOD FOR SOLVING GAS-DYNAMIC DIFFERENCE EQUATIONS*

Yu. P. POPOV and E. A. SAMARSKAYA

Moscow

(Received 19 February 1976)

THE CONVERGENCE conditions of Newton's iterative method applied to the solution of implicit difference schemes for one-dimensional non-stationary gas-dynamic equations in Lagrangian mass coordinates are investigated. The results obtained for the adiabatic case, taking into account linear pseudoviscosity, are compared with previously known conditions for isothermal flows in the absence of viscosity.

Iterative methods are usually used in gas-dynamic problems to solve implicit difference schemes, consisting of systems of non-linear algebraic equations. The numerical solution of a system of one-dimensional non-stationary gas-dynamic equations in Lagrangian mass coordinates by means of Newton's iterative method is described in [1, 2]. A theoretical analysis, and also the results of calculations, testify to the fact that in this case Newton's method possesses certain advantages over other iterative methods, for example the "explicit iteration" method, since it permits the use of coarse meshes with a comparatively large time step. Estimates of the convergence of Newton's iterative process were made in [1, 2] for the isothermal case without allowing for pseudoviscosity.

In the present paper these estimates are generalized for the adiabatic case in the presence of linear viscosity.

1. The system of equations of gas dynamics, describing the one-dimensional plane non-stationary flow of a gas in Lagrangian mass variables for the adiabatic case, can be written in the form [2, 3]

$$\begin{aligned} \frac{\partial v}{\partial t} &= -\frac{\partial g}{\partial s}, & \frac{\partial x}{\partial t} &= v, & \frac{\partial \eta}{\partial t} &= \frac{\partial v}{\partial s}, & \frac{\partial \varepsilon}{\partial t} &= -g \frac{\partial v}{\partial s}, \\ g &= p + \omega, & \omega &= \Omega \left(\eta, \frac{\partial v}{\partial s} \right), & p &= P(\eta, T), & \varepsilon &= E(\eta, T). \end{aligned}$$

**Zh. vychisl. Mat. mat. Fiz.*, 17, 1, 276–280, 1977.

Here t is the time, x is the Eulerian variable, η is the specific volume, $s, ds = \eta^{-1} dx$, is the Lagrangian mass variable, v, p, ϵ, T are respectively the velocity, pressure, internal energy and temperature of the gas, ω is the viscosity, g is the so-called total pressure, and the time derivative is Lagrangian. The last two relations in (1) are the thermodynamic equations of state.

Equations (1) are solved in some domain $\Omega = \{0 < s < M, t > 0\}$, on whose boundaries for $s = 0$ and $s = M$ boundary conditions are specified, for example, the laws of variation of velocity or pressure with time.

For definiteness we will consider an ideal gas with the equations of state

$$p = RT/\eta, \quad \epsilon = ap\eta, \quad a = 1/(\gamma - 1),$$

and with the linear viscosity

$$\omega = -\frac{v}{\eta} \frac{\partial v}{\partial s}, \quad \text{where} \quad v = \begin{cases} \bar{v}, & \partial v / \partial s < 0, \\ 0, & \partial v / \partial s \geq 0. \end{cases}$$

To construct a difference scheme approximating the system of differential equations (1), we introduce in the domain Ω a mesh, uniform for simplicity,

$$\omega_{h\tau} = \{(s_i, t_j), i=0, 1, \dots, N, j=0, 1, 2, \dots; s_{i+1} = s_i + h, t_{j+1} = t_j + \tau\}.$$

To the nodes (s_i, t_j) ("integral points") of the mesh $\omega_{h\tau}$ we refer the functions $x = x_i^j, v = v_i^j$,

and to the "half-integral points" $(s_{i+1/2}, t_j)$ we refer the functions $p = p_{i+1/2}^j, \eta = \eta_{i+1/2}^j, \epsilon = \epsilon_{i+1/2}^j, T = T_{i+1/2}^j, \omega = \omega_{i+1/2}^j, g = g_{i+1/2}^j$, where $s_{i+1/2} = s_i + h/2$.

The completely conservative difference scheme approximating the system of equations (1) in the case of an ideal gas, has the form [1, 2, 4]

$$\begin{aligned} v_i &= -g_s^{(\sigma)}, & x_i &= v^{(0.5)}, & \eta_i &= v_s^{(0.5)}, & \epsilon_i &= -g^{(\sigma)} \eta_i, \\ g &= p + \omega, & \omega &= -v v_s / \eta, & p &= RT/\eta, & \epsilon &= ap\eta. \end{aligned} \quad (2)$$

The parameter $0 \leq \sigma \leq 1$ is arbitrary.

The scheme (2) is written in the indexless notation [1, 2, 4, 5]

$$\begin{aligned} y &= y_i^j, & \hat{y} &= y_i^{j+1}, & y^{(\sigma)} &= \sigma \hat{y} + (1 - \sigma) y, \\ y_i &= (\hat{y} - y) / \tau, & y_s &= (y_{i+1} - y_i) / h, & y_s^- &= (y_i - y_{i-1}) / h. \end{aligned} \quad (3)$$

Transforming the last four equations in (2), we can rewrite this system of equations in the form

$$\begin{aligned} v_i &= -g_s^{(\sigma)}, & x_i &= v^{(0.5)}, & \eta_i &= v_s^{(0.5)}, & \epsilon_i &= -g^{(\sigma)} \eta_i, \\ \epsilon - ag\eta - avv_s &= 0. \end{aligned} \quad (2')$$

For $\sigma = 0$ the system of difference equations (2) or (2') is solved explicitly, but one thereby obtains a scheme conditionally stable for an extremely strict constraint on the time step of the mesh. In the acoustic approximation the stability condition has the form $\tau \leq kh^2$, where k is some constant [1, 2]. For $\sigma \geq 0.5$ the difference scheme (2) is unconditionally stable, but in this case iterative methods must be used to solve it.

2. The application of Newton's method to the system of non-linear equations (2) for finding the unknown values of the mesh functions $v, x, p, \eta, \varepsilon, T, \omega, g$ on the top time layer t_{j+1} for $\sigma \geq 0.5$ leads to the equations

$$\begin{aligned} \delta v + \sigma \tau \delta g_s &= -f_1, & \delta x - 0.5 \tau \delta v &= -f_2, & \delta x_s - \delta \eta &= -f_3, \\ \delta \varepsilon + g^{(0)} \delta \eta + \sigma \eta_s \delta g &= -f_4, & \delta \varepsilon - a \eta \delta g - a g \delta \eta - a v \delta v_s &= -f_5. \end{aligned} \quad (4)$$

Here δy is the difference in the values of the mesh function y in the adjacent $(k+1)$ -th and k -th intervals:

$$\delta y = \delta y^{[k+1]} = y^{[k+1]} - y^{[k]}. \quad (5)$$

Here all the unknown increments δy have the iteration number $k+1$, and the coefficients of the equations $g^{(0)}, \eta_s, g, \eta$ and the right sides $f_p, p=1, 2, \dots, 5$, are calculated at the lower k -th iteration and are regarded as known. As the initial "zeroth" iteration $y^{[0]}$ the preceding time layer $y^{[0]} = y^j$ can be used.

After the elimination of all the unknown functions, except δv , the system of linear equations (4) reduces at each mesh node to a three-point equation

$$A_i \delta v_{i-1} - C_i \delta v_i + B_i \delta v_{i+1} = -F_i, \quad i=1, 2, \dots, N-1, \quad (6)$$

whose coefficients A_i, B_i, C_i and the right side F_i depend only on the values of the mesh functions at the k -th iteration and j -th time layer.

Equations (6) are solved at each iteration by pivotal condensation [2, 5], the iterations are continued until some stability condition is satisfied, for example, the increments δv at all the mesh nodes become fairly small.

3. We study the convergence of the Newtonian iterative process described above. We subtract from each equation of the system (4) the corresponding equation of the system (2'). As a result we obtain the following system of linear equations:

$$\begin{aligned} \Delta v^{[k+1]} + \sigma \tau \Delta g_s^{[k+1]} &= 0, & \Delta x^{[k+1]} - 0.5 \tau \Delta v^{[k+1]} &= 0, \\ \Delta x_s^{[k+1]} - \Delta \eta^{[k+1]} &= 0, \\ \Delta \varepsilon^{[k+1]} + \sigma \eta^{[k]} \Delta g^{[k+1]} + \sigma g^{[k]} \Delta \eta^{[k+1]} - \sigma \Delta \eta^{[k]} \Delta g^{[k]} + \sigma \eta^j \Delta g^{[k+1]} & \\ + (1 - \sigma) g^j \Delta \eta^{[k+1]} &= 0, \\ \Delta \varepsilon^{[k+1]} - a \eta^{[k]} \Delta g^{[k+1]} - a g^{[k]} \Delta \eta^{[k+1]} + a \Delta \eta^{[k]} \Delta g^{[k]} - a v \Delta v_s^{[k+1]} &= 0. \end{aligned} \quad (7)$$

Here $\Delta y^{[k]} = y^{[k]} - y^{j+1}$ is the difference between the value of the mesh function at the k -th iteration and the exact solution of the difference problem. We note that this notation differs from the notation of (5), where the difference is taken between adjacent iterations.

The system (7) is more suitable for theoretical analysis. Eliminating from (7) all the functions $\Delta y^{[k+1]}$, except $\Delta g^{[k+1]}$, we arrive at the equation

$$A_i^{[k]} z_i^{[k+1]} - B_i^{[k]} (z_{i+1}^{[k+1]} + z_{i-1}^{[k+1]}) = F_i^{[k]}, \quad i=1, 2, \dots, N-1, \quad (8)$$

where

$$\begin{aligned} z_i^{[k]} &= \Delta g_i^{[k]}, \quad A_i^{[k]} = (\sigma + a) \eta_i^{[k]} - \sigma \eta_i^{j+1} + 2B_i^{[k]}, \\ B_i^{[k]} &= \sigma \frac{\tau}{h} \left[0.5 \frac{\tau}{h} ((\sigma + a) g_i^{[k]} + (1 - \sigma) g_i^{j+1}) + a v \right], \\ F_i^{[k]} &= (\sigma + a) \Delta \eta_i^{[k]} z_i^{[k]}. \end{aligned}$$

Obviously, $B_i^{[k]} > 0$. We require that the following inequality be satisfied:

$$D_i^{[k]} = A_i^{[k]} - 2B_i^{[k]} = (\sigma + a) \eta_i^{[k]} - \sigma \eta_i^{j+1} > 0. \quad (9)$$

We note that the inequality $A_i^{[k]} > 0$ is then also satisfied.

For definiteness we restrict ourselves to the problem in which on the boundaries of the domain Ω the variation of pressure with time is specified. Taking into account also the fact that

at the boundary points the pseudo-viscosity is assumed to be "zeroed" $\omega_0^{j+1} = \omega_N^{j+1} = 0$ (see [1]), we have for Eqs. (8) the boundary conditions $z_0^{[k+1]} = z_N^{[k+1]} = 0$. For the inhomogeneous equation (8) with the homogeneous boundary conditions to satisfy the inequality (9) we use the maximum principle [2, 5], which in particular implies that

$$\|z^{[k+1]}\|_C \leq \left\| \frac{F^{[k]}}{D^{[k]}} \right\|_C = \left\| \frac{(\sigma + a) \Delta \eta^{[k]} z^{[k]}}{(\sigma + a) \eta^{[k]} - \sigma \eta} \right\|_C \leq q_k \|z^{[k]}\|_C, \quad (10)$$

where

$$\|y\|_C = \max_i |y_i|, \quad q_k = \left\| \frac{(\sigma + a) (\eta^{[k]} - \hat{\eta})}{(\sigma + a) \eta^{[k]} - \sigma \eta} \right\|_C = \left\| \frac{\eta^{[k]} - \hat{\eta}}{\eta^{[k]} - \sigma \eta / (\sigma + a)} \right\|_C.$$

Obviously, Newton's iterative process converges if $q_k < 1$, and condition (9), which can be written in the form

$$\eta^{[k]} > \frac{\sigma}{\sigma + a} \eta. \quad (11)$$

is also satisfied. The inequality $q_k < 1$ is satisfied if for all $i = 1, 2, \dots, N$

$$-1 < \frac{\eta_i^{[k]} - \hat{\eta}_i}{\eta_i^{[k]} - \sigma \eta_i / (\sigma + a)} < 1. \quad (12)$$

holds. We consider two possibilities.

First, corresponding to the process of rarefaction of the gas, when $\hat{\eta} > \eta$. Then in (12) the condition on the right is satisfied automatically, and the one on the left leads to the inequality

$$\eta^{[k]} > \frac{1}{2} \left(\hat{\eta} + \frac{\sigma}{\sigma + a} \eta \right), \quad (13)$$

which simultaneously guarantees the satisfaction of condition (11).

In the contrary case ($\hat{\eta} < \eta$), corresponding to compression of the gas, condition (12) is satisfied if the inequality (13) holds, and also the inequality

$$\hat{\eta} > \frac{\sigma}{\sigma + a} \eta, \quad (14)$$

condition (11) being also satisfied.

It is reasonable to assume, and this is confirmed by calculations, that to satisfy the inequality (13) for any k it is sufficient to require its satisfaction at the zeroth iteration $k = 0$. Condition (13) for $k = 0$ ($\eta^{[0]} = \eta$) can be transformed to the form

$$\hat{\eta} - \eta < \frac{a}{\sigma + a} \eta,$$

whence after division by τ , using the notation of (3), and also the formula of difference differentiation $\rho_t = -\eta_t / \hat{\eta} \eta$, $\eta = 1/\rho$ (ρ is the density), we have

$$(1 + \sigma/a) \tau \hat{\eta} \rho_t > -1. \quad (15)$$

We note that for the case of compression $\rho_t > 0$, and thereby condition (15) is satisfied. After simple transformations the inequality (14) is reduced to the form

$$(1 + \sigma/a) \tau \hat{\eta} \rho_t < 1. \quad (16)$$

Combining (15) and (16), we arrive at the condition

$$(1 + \sigma/a) \tau \|\hat{\eta} \rho_t\|_C < 1. \quad (17)$$

We note that the condition for the convergence of the iterations obtained in [2, 4] for the isothermal case without allowing for pseudo-viscosity has the form

$$\tau \|\hat{\eta}\|_C \|\rho_t\|_C < 1. \quad (18)$$

Condition (17) for the isothermal case $\gamma=1$, $a=\infty$ becomes the inequality

$$\tau \|\hat{\eta}\rho_i\|_C < 1,$$

which is less "strict" than (18). We note that the presence of pseudo-viscosity in the scheme has no effect on the stability condition (17) or on the estimate of the rate of convergence (10). It must be pointed out that the convergence of the iterative process (10) investigated above has the nature of a geometrical progression with denominator q_k , while in [2, 4] the nature of the convergence studied for the isothermal case is quadratic. From the inequality (17) it also follows that in the adiabatic case the constraint on the mesh step is stricter than in the isothermal case.

Translated by J. Berry.

REFERENCES

1. POPOV, Yu. P. and SAMARSKII, A. A. Completely conservative difference schemes. *Zh. vychisl. Mat. mat. Fiz.*, 9, 4, 953-958, 1969.
2. SAMARSKII, A. A. and POPOV, Yu. P. *The difference schemes of gas dynamics* (Raznostnye skhemy gazovoi dinamiki), "Nauka", Moscow, 1975.
3. ROZHDESTVENSII, B. L. and YANENKO, N. N. *Systems of quasilinear equations* (Sistemy kvazilineinykh uravnenii), "Nauka", Moscow, 1968.
4. POPOV, Yu. P. and SAMARSKII, A. A. Methods for the numerical solution of one-dimensional non-stationary problems of gas dynamics. *Zh. vychisl. Mat. mat. Fiz.*, 16, 6, 1503-1518, 1976.
5. SAMARSKII, A. A. *Introduction to the theory of difference schemes* (Vvedenie v teoriyu raznostnykh skhem). "Nauka", Moscow, 1971.

BOOK REVIEW*

K. SARKHADI and I. VINCZE. *Mathematical methods of statistical quality control*. 415p.
Akademiai Kiado, Budapest, 1974.

Statistical control of the quality of production is one of the most important fields of application of the theory of probability and mathematical statistics. There are a number of books (including some in Russian), devoted to this field of applied mathematics, but many of them are at too low a mathematical level, and hence they do not satisfy the demands of modern complex production.

In the book reviewed, written by two prominent Hungarian specialists, the methods of mathematical quality control are based on the serious foundation of the theory of probability and mathematical statistics. Therefore a large part of the book is devoted to the principles of these disciplines, without forgetting the main purpose of the book.

Of course it will not serve as a textbook on the theory of probability and mathematical statistics, since it expounds the principles of these sciences in outline, without detailed proofs of the theorems, without a large number of training examples and problems, which are characteristic of every textbook. Accordingly the book is intended for a wide range of applied mathematicians, engineers, students of higher educational establishments and others interested in the applications of the theory of probability and mathematical statistics. For all these categories of readers it will constitute an excellent reference book.

The book consists of three parts and an appendix. The first part is introductory. Here the fundamental concepts and definitions used in the following parts are explained. The second part occupies more than half the book. Here the fundamental theorems of probability are explained. In particular the fundamental types of probability distribution, the principles of the sampling method, theories of order statistics etc. are explained in great detail. Also explained are the fundamental methods of mathematical statistics: the theory of estimation, the statistical testing of hypotheses, the principles of correlation and regression analysis, the theory of statistical decisions, the principles of stochastic processes etc. The third part is devoted to methods of statistical quality control. Here, after presenting the fundamental ideas, detailed explanations are given of the methods of continuous control and acceptance control, based on the main types of control charts and on various standards. The last paragraph of this part is devoted to the principles of reliability theory.

The book is supplemented by 14 basic statistical tables. Among them there are tables of random numbers, the normal probability density and function, the Poisson distribution, the binomial distribution, the percentage points of the F -distribution, Student's distribution, the χ^2 -distribution, the Kolmogorov-Smirnov distribution etc. There is an extensive bibliography, separated into books, papers, tables and standards, and also the necessary indexes. The printing of the book is excellent. The translation of this fundamental work into Russian is extremely desirable.

M. K. Kerimov

Translated by J. Berry.

*Zh. vychisl. Mat. mat. Fiz., 17, 1, 281, 1977.